DECE Volume 93, Number 1, February 1985

Merton H. Miller and Charles W. Upton: A Test of the Hotelling Valuation Principle

Clas Bergström, Glenn C. Loury, and Mats Persson: Embargo Threats and the Management of Emergency Reserves

Torsten Persson and Lars E. O. Svensson: Current Account Dynamics and the Terms of Trade: Harberger-Laursen-Metzler Two Generations Later

John P. Gould and Robert E. Verrecchia: The Information Content of Specialist Pricing

Yang-Ming Chang and Isaac Ehrlich: On the Economics of Compliance with the Minimum Wage Law

Gur Huberman and G. William Schwert: Information Aggregation, Inflation, and the Pricing of Indexed Bonds

David Cox and Richard Harris: Trade Liberalization and Industrial Organization: Some Estimates for Canada

Scott Freeman: Transactions Costs and the Optimal Quantity of Money

William R. Johnson: The Economics of Copying

Paul A. Spindt: Money Is What Money Does: Monetary Aggregation and the Equation of Exchange

Daniel S. Hamermesh and James M. Johannes: Food Stamps as Money: The Macroeconomics of a Transfer Program

The University of Chicago Press

Volume 93, Number 1, February 1985

Merton H. Miller and Charles W. Upton: A Test of the Hotelling Valuation Principle

Clas Bergström, Glenn C. Loury, and Mats Persson: Embargo Threats and the Management of Emergency Reserves

Torsten Persson and Lars E. O. Svensson: Current Account Dynamics and the Terms of Trade: Harberger-Laursen-Metzler Two Generations Later

John P. Gould and Robert E. Verrecchia: The Information Content of Specialist Pricing

Yang-Ming Chang and Isaac Ehrlich: On the Economics of Compliance with the Minimum Wage Law

Gur Huberman and G. William Schwert: Information Aggregation, Inflation, and the Pricing of Indexed Bonds

David Cox and Richard Harris: Trade Liberalization and Industrial Organization: Some Estimates for Canada

Scott Freeman: Transactions Costs and the Optimal Quantity of Money

William R. Johnson: The Economics of Copying

Paul A. Spindt: Money Is What Money Does: Monetary Aggregation and the Equation of Exchange

Daniel S. Hamermesh and James M. Johannes: Food Stamps as Money: The Macroeconomics of a Transfer Program

The University of Chicago Press

lation is a constant different from zero. Thus, for example, one might attempt to hedge an inventory of cash good by *purchasing* futures for a good which is a complement in the sense that $\sigma_{01} < 0$.

The generalization to multiple futures contracts says that the proportion of output that should be hedged in each contract is given by the coefficient of the theoretical multiple regression of cash prices on the n futures prices ($\Sigma_{11}^{-1}\Sigma_{10}$ is the vector of these multiple regression coefficients). This result suggests that the partial (given remaining futures prices) correlation coefficient between the cash price and a futures contract is a better indicator of the usefulness of that contract for hedging purposes than is the simple correlation coefficient. Thus it is conceivable that the optimal hedge against a *long* cash position involves a *long* futures position in a contract even when the (simple) correlation between the cash and that futures price is positive.

Finally, it is worthwhile pointing out that the pure hedge term, $y\Sigma_{11}^{-1}\Sigma_{10}$, is equal to the risk-minimizing futures position corresponding to a *predetermined* cash position y. A hedger thus uses the possibilities offered by futures markets to minimize his risk, then he speculates. This interpretation should not, however, conceal that futures and cash positions are *jointly* determined and that the risk-minimizing futures position is generally suboptimal since it ignores the opportunities for speculative return and because it is decided conditionally on a cash position that will generally be suboptimal. It should be clear from the previous analysis that the larger $|\sigma_{01}/\sigma_{11}|$, the more effective in reducing risk is a single cross hedge. This criterion could be used for selecting which delivery month, which exchange, or even which underlying commodity would offer the best risk-reducing opportunity.

IV. The Optimal Cash Position

Equation (5) characterizes the optimal futures position, given y; inserting (5) into (3) gives the necessary condition for y given that f is adjusted optimally as y varies. After simplification, this condition is

$$\bar{p}_1 + \Sigma_{01}\Sigma_{11}^{-1}(p_0^f - \bar{p}_1^f) - y\chi(\sigma_{00} - \Sigma_{01}\Sigma_{11}^{-1}\Sigma_{10}) - c'(y) = 0. \qquad (7)$$

Equation (7) is nothing else than an equalization of marginal revenue and marginal cost. The latter is the sum of the two last terms: $c'(y)$ is the marginal cost of production; $\chi y(\sigma_{00} - \Sigma_{01}\Sigma_{11}^{-1}\Sigma_{10})$ is the marginal risk premium. The risk premium can be rewritten as $\chi y\sigma_{00}(1 - R^2)$, where R^2 is the theoretical multiple correlation coefficient squared. Thus the risk premium depends on the degree of risk aversion, the size of the cash position, the variability of cash prices, and the proportion of that variability which cannot be eliminated through the opti

mal hedge in the n futures contracts. For a given cash price variance, the risk premium decreases as the multiple correlation coefficient increases; consequently, the higher the R^2, the larger the cash position (in absolute value).

The two first terms of (7) can be viewed as a "planning price," used as a substitute for the as yet unknown cash price. The planning price is a weighted sum of the expected cash price and *all* current and next-period futures prices. The weight given each futures price depends on its coefficient in the multiple regression and thus corresponds to the proportion of the cash position hedged on that particular market. Again it is enlightening to look at the simple cross-hedge case ($n = 1$); the planning price can then be written

$$(1 - \beta)\bar{p}_1 + \beta(p_0^f - \bar{B}), \tag{8}$$

where $\beta = \sigma_{01}/\sigma_{11}$ and $\bar{B} = \bar{p}_1^f - \bar{p}_1$ is the expected value of what is known as the basis, the difference between futures and cash price. The appropriate allocation signal for the decision maker is now a linear combination of the expected cash price and the currently quoted futures prices "scaled" for the expected difference between futures and cash prices. The interest of this formulation is in the fact that people in the trade often assert that bases (i.e., price differences) are more easily predictable than price levels. Equation (8) shows that the closer the relationship between cash and futures, the more the hedger can rely on futures price and basis expectation rather than price expectation for "pricing" his output.

From this analysis, it is quite clear that an increase in the absolute value of the covariance between futures and cash prices, $|\sigma_{01}|$, may have conflicting effects on real production. On the one hand, it decreases the riskiness associated with the "real" activity of the producer; hence the proportion of marginal revenue allocated to risk remuneration decreases and the cash position can be made larger in absolute value. On the other hand, the heavier reliance on the futures price in making production decisions may entail a smaller cash position depending on its level relative to the expected cash price.

In the general case ($n > 1$), we can also interpret the planning price in terms of the vector of bases, of the form $p_{it}^f - p_t$. For the first two terms of (7) can be written as

$$(1 - \beta\iota)\bar{p}_1 + \beta(p_0^f - \bar{B}_1), \tag{9}$$

where $\beta = \Sigma_{01}\Sigma_{11}^{-1}$, ι is the unit vector, and $\bar{B}_1 = \bar{p}_1^f - \bar{p}_1\iota$ (the vector of the expected bases).[4] Thus observed futures prices and basis ex-

[4] We could also conceive of the following alternative generalization of basis for the case of $n > 1$. Let the generalized basis Bg be $Bg = \Sigma_{01}\Sigma_{11}^{-1}p_1^f - p_1$; this is the difference of

Journal of Political Economy

Volume 89 Number 3 June 1981

Articles

429 The Deterrent Effect of Antitrust Enforcement
Michael Kent Block, Frederick Carl Nold, and Joseph Gregory Sidak

446 The Market Evaluation of Human Capital: The Case of Indentured Servitude
David W. Galenson

468 The Welfare Cost of Capital Income Taxation in a Growing Economy
Christophe Chamley

497 Foreign Ownership and the Theory of Trade and Welfare
Richard A. Brecher and Jagdish N. Bhagwati

512 Accounting for Price Changes: American Steel Rails, 1879–1910
Robert C. Allen

529 Some Evidence on Cross-Sector Effects of the Minimum Wage
George E. Tauchen

548 Taxes, Inflation, and the Durability of Capital
Andrew B. Abel

561 Uncertain Lifetime, Consumption, and Dissaving in Retirement
James B. Davies

Review Article

578 What Kind of a Science Is Economics? A Review Article on *Causality in Economics* by John R. Hicks
Christopher A. Sims

Comments

584 To Save or Savor: The Rate of Return to Storing Wine
Elizabeth Jaeger

593 On the Relationship between Commodity Price Changes and Factor Owners' Real Positions
James Cassing

Confirmations and Contradictions

596 Stochastic Implications of the Life Cycle–Permanent Income Hypothesis: Evidence for the U.K. Economy
Vince Daly and George Hadjimatheou

Miscellany

600 Why There Are No Risk Preferrers
David Friedman

Book Reviews

601 Ronald G. Ehrenberg, *The Regulatory Process and Labor Earnings*
Orley Ashenfelter

603 Friedrich A. Hayek, *Law, Legislation and Liberty. A New Statement of the Liberal Principles of Justice and Political Economy*. Vol. 3: *The Political Order of a Free People*
Clas Wihlborg

609 Steven A. Lippman and John J. McCall, eds., *Studies in the Economics of Search*
Dale T. Mortensen

611 Harold Lydall, *A Theory of Income Distribution*
Martin Bronfenbrenner

Journal of Political Economy

Volume 89 Number 4 August 1981

Articles

615 The Role of Market Forces in Assuring Contractual Performance
Benjamin Klein and Keith B. Leffler

642 The Political Economy of Benefits and Costs: A Neoclassical Approach to Distributive Politics
Barry R. Weingast, Kenneth A. Shepsle, and Christopher Johnsen

665 Flexible Exchange Rates, Prices, and the Role of "News": Lessons from the 1970s
Jacob A. Frenkel

706 The Role of Intergenerational Transfers in Aggregate Capital Accumulation
Laurence J. Kotlikoff and Lawrence H. Summers

733 Output Variability under Monetary Policy and Exchange Rate Rules
Warren E. Weber

752 Point Spreads versus Odds
Gilbert W. Bassett, Jr.

769 Time Preference and International Lending and Borrowing in an Overlapping-Generations Model
Willem H. Buiter

798 Mobility Costs, Frictional Unemployment, and Efficiency
Peter A. Diamond

Comments

813 International Transmission of a Real Shock under Flexible Exchange Rates: A Comment
Betty C. Daniel

819 Market Constraints as a Rationale for the Friedman-Savage Utility Function
Elie Appelbaum and Eliakim Katz

Book Reviews

826 Jagdish N. Bhagwati, *Foreign Trade Regimes and Economic Development: Anatomy and Consequences of Exchange Control Regimes*
Deepak Lal

830 Richard A. Easterlin, ed., *Population and Economic Change in Developing Countries*
John C. Caldwell

834 Paul W. MacAvoy, *The Regulated Industries and the Economy*
Wesley A. Magat

837 Samuel L. Popkin, *The Rational Peasant: The Political Economy of Rural Society in Vietnam*, and Alan Macfarlane, *The Origins of English Individualism: The Family, Property and Social Transition*
Donald N. McCloskey

Journal of Political Economy

Volume 89 Number 5 October 1981

Articles

841 Rank-Order Tournaments as Optimum Labor Contracts
Edward P. Lazear and Sherwin Rosen

865 An Exploration in the Theory of Exchange-Rate Regimes
Elhanan Helpman

891 Subsidies to New Energy Sources: Do They Add to Energy Stocks?
William J. Baumol and Edward N. Wolff

914 A Rational Theory of the Size of Government
Allan H. Meltzer and Scott F. Richard

928 The Family, Inheritance, and the Intergenerational Transmission of Inequality
Nigel Tomes

959 Intraindustry Specialization and the Gains from Trade
Paul R. Krugman

974 The Adjustment of Consumption to Changing Expectations about Future Income
Marjorie A. Flavin

1010 Measurement of Monopoly Behavior: An Application to the Cigarette Industry
Daniel A. Sumner

1020 Swedish Tax Rates, Labor Supply, and Tax Revenues
Charles E. Stuart

Comments

1039 A Superior Solution to Captain MacWhirr's Problem: An Illustration of Information Problems and Entitlement Structures
Gene E. Mumy

1044 Concentration Changes and Inflation: Some Evidence
E. Woodrow Eckard, Jr.

Book Reviews

1052 Sir John R. Hicks, *The Crisis in Keynesian Economics*
Karl Brunner

1054 Lester C. Thurow, *The Zero-Sum Society: Distribution and the Possibilities for Economic Change*
Wilfred Beckerman

Journal of Political Economy

Volume 89 Number 6 December 1981

Articles

1059 An Empirical Model of Labor Supply in a Life-Cycle Setting
Thomas E. MaCurdy

1086 Output Effects of Government Purchases
Robert J. Barro

1122 Information Costs, Duration of Search, and Turnover: Theory and Applications
Louis L. Wilde

1142 Macroeconomic Policy, Exchange-Rate Dynamics, and Optimal Asset Accumulation
Maurice Obstfeld

1162 Wage and Employment Determination under Trade Unionism: The International Typographical Union
James N. Dertouzos and John H. Pencavel

1182 Cross Hedging
Ronald W. Anderson and Jean-Pierre Danthine

1197 The Implications of Competition among Jurisdictions: Does Tiebout Need Politics?
Dennis Epple and Allan Zelenitz

1218 Monetary Expansion and Real Exchange Rate Dynamics
Nissan Liviatan

1228 Economies of Scale and Barriers to Entry
Richard Schmalensee

1239 The Great Depression and Commodity-exporting LDCs: The Case of Brazil
Eliana A. Cardoso

Confirmations and Contradictions

1251 Estimating Property Tax Capitalization: A Further Comment
Raymond M. Reinhard

1261 Tax Changes and Cigarette Prices
Michael T. Sumner and Robert Ward

Miscellany

1266 'Twas a Night in the Sixties
Martin S. Feldstein

1270 Policy Research
J. Michael Finger

1275 Index to Volume 89

Positive Time Preference

Mancur Olson and Martin J. Bailey

University of Maryland

The case for positive time preference is absolutely compelling, unless there is an infinite time horizon with the expectation of unending technological advance combined with what we call "drastically diminishing marginal utility." This finding holds both in the positive and normative senses. A corollary is that savings are interest elastic.

Most writers on capital theory have taken it for granted that most people systematically prefer present to future consumption. This positive rate of time preference is usually ascribed to a "faulty telescopic faculty," to "myopia," or to "impatience," even though this implies an irrationality or continuing perceptual bias that is inconsistent with the behavioral assumptions economists usually use. To our knowledge no one has ever provided convincing evidence that there is in fact normally positive time preference, or even specified an empirical test capable of determining whether there is or not. Apparently the belief in myopia often rests on subjective judgments about the relative utility levels of other individuals across different time periods that have very limited scientific value. In other cases a preference for present over future consumption is erroneously supported with evidence that, as we shall argue, need reflect nothing more than a rational adaptation to positive interest rates generated by the fact that capital goods can be useful in production. In still other instances time preference is supposed to follow directly from the fact that there is

We are indebted to Gary Becker, Christopher Clague, Donald Dewey, Robert Dorfman, Tjalling Koopmans, Peter Murrell, George Stigler, Paul Wonnacott, and Howell Zee for helpful comments and suggestions. Part of the work on this article was supported by grants from the Environmental Protection Agency, the National Science Foundation, and Resources for the Future.

[*Journal of Political Economy*, 1981, vol. 89, no. 1]
© 1981 by The University of Chicago. 0022-3808/81/8901-0007$01.50

always uncertainty about the future, so the consumer cannot be sure whether any assets that he acquires by saving will retain their value or even that he will be alive to enjoy any fruits of his saving. In fact, uncertainty about future economic conditions and the duration of one's life can increase as well as decrease the incentive to save.

This paper develops an approach to intertemporal choice that makes it possible to determine from actual household choices in certain circumstances whether or not there is positive time preference. The empirical tests on ordinary saving and consumption behavior that grow out of this approach suggest that a decidedly positive rate of time preference is typical if not almost universal. There is an alternative interpretation of the evidence that permits doubts in cases where the planning horizon is infinite, and this we label "drastically diminishing marginal utility." This intriguing phenomenon could in principle be widespread. In long-run contexts it is so similar to time preference in its implications that the two are difficult to distinguish. Nonetheless, especially when short- as well as long-run contexts are considered, the case that there is often time preference becomes compelling, even if drastically diminishing marginal utility should also be present.

We shall also show that those who advocate certain environmental and resource policies on the ground that the interests of all future years and generations ought to be weighted equally with our concern for utility in the present are contradicted by their own behavior. We will similarly show that it is logically inconsistent to argue that there is positive time preference and that household savings do not generally respond to changes in interest rates.

I. Positive Time Preference and Diminishing Marginal Utility

Often the idea of time preference appears in an impressionistic form without any workable definitions (see, e.g., Pigou 1960; Jevons 1965), though there are also a number of quite useful efforts to obtain a precise understanding of the matter.[1] Interestingly enough, the clearest conception of positive time preference that we have been able to find was in Bohm-Bawerk's original account.

> *Present goods have in general greater subjective value than future (and intermediate) goods of equal quantity and quality . . .*
>
> A *first* principal cause capable of producing a difference in

[1] See the helpful discussion in Marglin (1963), Feldstein (1964), and Hirshleifer (1970). There are also important articles that appeared somewhat earlier, notably Samuelson (1937), Strotz (1956), Debreu (1959), and Koopmans (1960).

> value between present and future goods is . . . if a person suffers in the present from appreciable lack of certain goods, or of goods in general, but has reason to hope to be more generously provided for at a future time, then that person will always place a higher value on a given quantity of immediately available goods than on the same quantity of future goods. This situation occurs with very great frequency in our economic life. . . .
>
> We must now consider a *second* phenomenon of human experience—one that is heavily fraught with consequence That is the fact that we feel less concerned about future sensations of joy and sorrow simply because they do lie in the future, and the lessening of our concern is in proportion to the remoteness of that future. Consequently we accord to goods which are intended to serve future ends a value which falls short of the true intensity of their future marginal utility. [1959, pp. 265–73]

This quotation suggests two distinct elements, each of which might influence intertemporal choice: (1) diminishing marginal utility of consumption at any given time, where utility is an unchanging function of the amount of consumption at the time, and (2) discounting of future versus present utility. This distinction is so fundamental that we believe there can be no adequate explanation of intertemporal choice without it, yet it is too often ignored Each of these two elements, moreover, naturally suggests a testable hypothesis that can be proven false. Bailey, Olson, and Wonnacott have shown elsewhere that "The Marginal Utility of Income Does Not Increase" (1980) and in at least a great majority of cases is actually decreasing. In this paper we accordingly accept Bohm-Bawerk's assumption that there is diminishing marginal utility to consumption in each period and go on to explore whether the available evidence is consistent with the hypothesis that utility in the future is valued less than utility in the present. In order to do this, and to obtain the full range of hypotheses that can be drawn with Böhm-Bawerk's distinction, we shall first develop a more formal statement of the distinction.

Following a suggestion by Hirshleifer (1970),[2] we represent utility at moment t by a function $v(C_t)$; the consumption C_t is time subscripted, but the function is time invariant, depending only on the level or quantity of C_t, regardless of t. Then with time measured in terms of discrete periods of equal length, we write the consumer's

[2] The Hirshleifer analysis in turn has precedents in the articles by Samuelson (1937), Strotz (1956), Debreu (1959), and Koopmans (1960) cited in the preceding footnote

utility function to be maximized at time zero, subject to a wealth budget constraint, as

$$U = \sum_{t=0}^{T} \frac{v(C_t)}{(1+\eta)^t}, \tag{1}$$

where the consumer's time horizon for planning is T, which need not be finite, and where η is a constant, the rate per period of positive time preference. We assume that the first derivative v' is positive and (because of the finding mentioned before) that the second is negative, $v'' < 0$.

Now consider a household choosing among alternative combinations (C_0, C_1) of consumption at $t = 0$ and $t = 1$, holding its consumption pattern for all later periods fixed. If we take the differential of (1), setting $dC_t = 0$ for all $t \geq 2$, we have

$$dU = v'(C_0)dC_0 + \frac{v'(C_1)dC_1}{1+\eta}, \tag{2}$$

and setting $dU = 0$ to represent holding the household indifferent,

$$v'(C_0)dC_0 + \frac{v'(C_1)dC_1}{1+\eta} = 0,$$

or

$$\left(\frac{dC_1}{dC_0}\right)_{U = \text{const}} = -\frac{v'(C_0)}{v'(C_1)}(1+\eta) \tag{3}$$

The household will demand an increase in C_1 greater than the reduction in C_0, that is,

$$\left|\frac{dC_1}{dC_0}\right| > 1,$$

to hold utility constant, if either of two things is true

$$C_1 > C_0 \text{ with } \eta = 0, \tag{4}$$

or

$$\eta > 0 \qquad \text{with } C_1 = C_0. \tag{5}$$

Conditions (4) and (5) succinctly state the two causes of interest in the quoted passage of Bohm-Bawerk.

Inequality (4) follows because $v'' < 0$. Diminishing marginal utility of income implies that, if all else is equal, a household with greater consumption potential in the later period will be willing to borrow at a positive interest rate in order to make consumption in the two periods more nearly equal. This is the "first cause" of the preference

for present over future consumption mentioned in the quotation from Böhm-Bawerk.

The hypothesized second cause is represented by the constant η in the intertemporal utility function (1). When it is positive rather than zero as in (5), the household would be willing to pay some positive interest rate to increase consumption in the initial period at the expense of consumption in the later period, even with equal endowments in the two periods.

By tracing out the combinations of C_0 and C_1 to which the household is indifferent, using (1) or (3), assuming $v(C_t)$ and η are known, one can obtain an indifference curve between present and next-period consumption. That is, equation (1) is an implicit function of C_0 and C_1, when all other C_t are held constant and when U is also held constant, this function,

$$U_0 = v(C_0) + \frac{v(C_1)}{1 + \eta} + \text{const.},$$

is the equation of an indifference curve in (C_0, C_1) space. Such an indifference curve U_0 is shown in figure 1. Along the 45° line $C_0 = C_1$, and on this line any tangent to an indifference curve, such as ABD, has a slope of $-(1 + \eta)$. This can be seen by setting $C_0 = C_1$ in equation (3), or by noting that when consumption is the same in both periods, any preference for present over future consumption must be due to η.

Though the marginal rate of substitution between consumption in any two periods for a household in equilibrium is frequently called the rate of "time preference," which is of course equal to the interest rate, this casual usage has caused a lot of confusion. We therefore strongly suggest that "positive time preference" be defined to exclude the effect of a difference in marginal utility due to any lower level of consumption in the present and, therefore, to include only any preference for present over future consumption due to other causes. In our formulation, time preference is then defined as a positive value of the constant η, rather than in terms of the entire expression (3), which gives the slope of the indifference curve. The ratio of marginal utilities, $[v'(C_0)]/[v'(C_1)]$, depends not on time but on levels of consumption in the two periods; any preference for present over future utility, by contrast, can be conveniently represented by the subjective discount rate η.

II. Separable Utility

Before we can properly derive our positive results, we need to consider whether the view of intertemporal choice set out in equation (1)

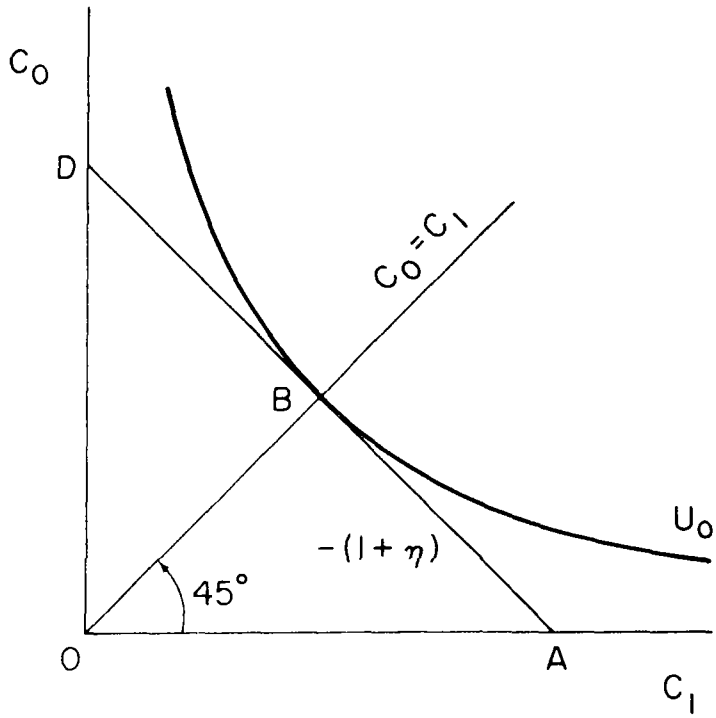


FIG. 1

is appropriate. In particular, we need to consider whether it is necessary or expedient to write utility as a function of consumption in every period.

$$U = U(C_0, C_1, C_2, \ldots, C_T). \tag{6}$$

This formulation, unlike that in equation (1), would formally encompass any complementarity between the levels of consumption in different periods. Indeed, one could go further and write utility as a function of the amount of each separate good consumed in each period—as would be necessary, for example, if the typical household consumed a little cream today to go with the coffee it will consume a week from Monday. In other words, the utility derived from a given level and pattern of consumption in a particular period could conceivably depend not only on the level but also on the composition of consumption in prior periods. There are examples of interdependence of consumption on different dates, such as a preference to have different dinner menus on successive days, and it could be necessary for some special purposes to take these into account. However, as

Samuelson (1937) has argued, functions that allow unlimited interrelationships become so general as to be almost vacuous. If a utility function rules nothing out, it will be consistent with almost any possible type of behavior and thus generate no potentially falsifiable hypotheses of interest. In addition, it has been shown elsewhere (Bailey and Olson 1977; Bailey et al. 1980) that the assumption of separability is more realistic than is usually supposed. Finally, it is hard to imagine how any plausible type of temporal complementarity could generate results that are qualitatively different from those in this article, with one obvious exception we can handle now.

An individual cannot, of course, derive utility from consumption in any period unless he has survived prior periods and also preserved the health needed to enjoy consumption. This could be accommodated formally by treating the essentials needed to preserve life and health as crucial investments in human capital. This would not really solve the problem, though, since the temporal interdependence would still have to be taken into account in investment decisions. Similarly, if a family is to have grandchildren or more distant descendants, then each intermediate generation must have at least enough resources to survive long enough to raise children. Thus we emphasize that our analysis must be modified before it can be applied in any situations where levels of consumption too low for healthful viability or reproduction are at issue. Abstractly, we deal with this starvation level of consumption ${}_0C$ by supposing that $v'(C) \to \infty$ as $C \to {}_0C^+$ and that $v(C)$ is undefined for $C < {}_0C$. Given this qualification, we can safely continue to use function (1).

III. Erroneous Inference of Myopic Behavior

In Section I we distinguished time preference from the effects of diminishing marginal utility of consumption in each period. With the aid of this distinction we can clear up one elementary yet surprisingly common error. Since Bohm-Bawerk's time, if not before, some economists have assumed that positive interest rates indicate that most people have positive time preference. Though we shall later provide evidence that time preference is in fact positive, this finding by no means follows from the existence of interest.

This is obvious once one considers the implications of the fact that capital goods can be useful in production. The existence of profitable investment opportunities implies that reduction of consumption by a dollar now will allow more than a dollar's increase in consumption in the future. A glance at equation (1) reminds us that even if time preference (η) were zero, this fact can generate positive interest rates, because diminishing marginal utility of consumption in each period

entails that consumers will accept a lower level of consumption in one period than another only if they are rewarded for doing so. Obviously, positive interest rates and the excess of marginal utility of present over that of future consumption that they induce[3] does not by itself justify the conclusion that there is any preference for present over future utility, much less persistent myopia or irrationality.

When one looks at both the demand and supply sides, it becomes clear that each of the following combinations of conditions is sufficient to generate the customary positive interest rates: (1) diminishing marginal utility plus profitable investment opportunities with nondecreasing income (and zero time preference); (2) diminishing marginal utility with increasing endowment income (and zero time preference), (3) positive time preference with constant income or with profitable investment opportunities (and nonincreasing marginal utility). Therefore, the fact that interest rates are positive does not by itself tell us which causes are operative. Only evidence of a different kind could demonstrate that there is positive time preference.

The rudimentary considerations just mentioned are, however, sufficient to tell us something about the relationship of η, whether it be positive, negative, or zero, and the prevailing interest rate. Consider initially a manna-from heaven world with no producers. Such a world could have a consumption-loans market with a determinate rate of interest in long-run equilibrium. Since we have not specified the times at which manna falls from heaven, or yet offered any evidence that would either entail or rule out positive time preference, the interest rate in this consumption-loans market could be anything—even zero or negative. If the manna falls evenly over time, the interest rate in this consumption-loans market is bound to be equal to η.

But now suppose that advantageous production processes are discovered in which produced capital is an argument in the production function for consumption goods. Since this discovery will increase the demand for borrowings but not shift the supply function for loans, the interest rate must rise. The result is that in equilibrium $r > \eta$. Except for a few extreme cases, the real world will be one in which $r > \eta$, because in almost all real-world cases capital goods have a positive net marginal product. The marginal rate of substitution between present and future consumption in long-run equilibrium will of

[3] The difficulties involved in any effort to get one unambiguous measure of the capital stock and the extended controversies that have grown out of this should not, of course, obscure the fact that capital accumulation requires postponement of consumption. Böhm-Bawerk's writings, despite their well known shortcomings, are sufficient to show this (see Dorfman 1959*a*, 1959*b*). Thus the presence of "reswitching," e.g., does not undermine our analysis here.

course always equal the rate of interest, but η is not in general equal to this marginal rate of substitution, even if η is positive.

IV. Is Time Preference Observable?

It is intuitively obvious from the discussion of figure 1 and from what has just been said that we can ascertain an individual's rate of time preference by observing the interest rate at which he chooses a level consumption stream. This means that we have a firm and simple basis for using a revealed preference approach to obtain solid scientific evidence about time preference and have no need to appeal to anything resembling introspection or interpersonal comparisons of utility.[4]

The conclusion that the household chooses a level consumption stream if the rate of interest equals the rate of pure time preference clearly holds for the multiperiod as well as for the two-period cases (see Samuelson 1937). This finding is itself enough to make time preference observable. We noted earlier that, unless incomes are declining, the existence of productive investment opportunities implies that the equilibrium rate of interest will be greater than the rate of time preference, and this is true however high the rate of time preference might be. Therefore we would predict that in most cases households would have positive levels of savings that generate a rising level of consumption over time. This theoretical prediction appears to be relatively realistic.

There are also other ways to measure time preference from revealed preferences. If the household maximizes utility (1) subject to the budget constraint

$$\sum_{t=0}^{T} \frac{Y_t^* - C_t}{(1+r)^t} = 0, \tag{7}$$

varying each C_t to obtain this maximum, the result implies

$$\frac{v'(C_t)}{v'(C_0)} = \frac{(1+\eta)^t}{(1+r)^t} \tag{8}$$

[4] In seeking revealed preference we are, of course, only continuing a line of inquiry started by others. In 1937 Samuelson proposed a hypothetical method for obtaining the *utility function* by integration, i.e., by mapping observed choices into a functional representation; in effect this was applying the revealed-preference approach to our problem. Similarly, Debreu (1959) showed, with a rigorous proof, that a "preordering," or a consistent, transitive set of preferences, implies a utility function that is unique up to a linear transformation. Frisch's (1959) analysis also proceeded in a similar spirit. There are also others who have made significant contributions to this way of researching the problem, such as Strotz (1956) and Hirshleifer (1970), whose contributions will be discussed later.

with a constant interest rate $\bar{r}$ for all t. Evidently, as we noted above, $C_t = C_0$ when $r = \eta$, so that when we find the rate of interest r_0 at which the household chooses a level consumption stream (i.e., zero savings from t_0), then $\eta = r_0$, and the value of η has been revealed.

Similarly, if $r > \eta$, as it usually is, $C_t > C_0$. If one has independent information on how rapidly $v'(C)$ declines with increasing C (see Bailey et al. 1980), one can estimate η simply by examining the chosen growth path of consumption and finding the η that reconciles it with equation (8).

V. Uncertainty

Though there are some notable exceptions, such as Ramsey (1928), and more recently Stigler and Becker (1977), most economists appear to take it for granted that time preference is positive. Sometimes this conclusion seems to rest not so much on the erroneous inferences we described in Section III as on the notion that the individual is uncertain about the date of his own death and will therefore discount prospective utility because of the possibility he will die earlier than expected. This argument is quite unsatisfactory for at least two reasons.

First, uncertainty about the date of one's death can increase as well as decrease an individual's provision for the future, because it can create an incentive to save more to provide undiminished consumption in the event of an unexpectedly long life. If income could costlessly be converted into an annuity, this would not be the case, but ideal annuities are not available, and in less developed economies they often cannot be purchased at all. It is an empirical question whether uncertainty about the length of one's life increases or decreases provision for the future.

A second reason why uncertainty about the date of one's death does not entail positive time preference is evident from the fact that there is normally positive saving in the society as a whole. Even if there were negative time preference over the life cycle, that is, a preference for utility in later as against earlier years of life, a society with no population growth or technological advance would have no social saving whatever if everyone consumed his entire output over the course of life: The savings in one set of years would just equal dissaving in another. In fact, the developed societies have saved enough to provide for a considerable long-term growth of tangible capital per workers as well as a vast increase in the amount of education and human capital. This substantial social saving had to be due in large part to a desire to bequeath increased consumption opportunities to descendants. The desire to bequeath assets to descendants provides

another reason why the argument that uncertainty about the date of one's death implies positive time preference is inadequate.

There is also uncertainty about whether the assets obtained through savings will generate the promised yield or even preserve their value. Such risks lower certainty-equivalent rates of return, and so are different and distinct from time preference. Risks may be reduced through diversification of portfolios, but transactions costs limit the extent to which risk can be eliminated in this way for the household of average means, and there are in addition great problems for such a household in diversifying against the risks to property in any single country (or single set of countries that might have a common destiny). Thus the risk that investments will prove unrewarding would appear to reduce saving, just as a positive rate of pure time preference would.

Uncertainty nevertheless also has two important countervailing effects that have been neglected, if noticed at all, in the literature. The first of these arises from risk aversion and from the absence of futures markets and insurance markets for most goods. The absence of futures markets makes it impossible for a typical household to hedge against a loss of demand for whatever specialized skill it has to offer, or against the loss in value of whatever other assets it possesses, or against a depression. Similarly, a typical household cannot insure against the birth of a child that will never be capable of supporting itself or even fully insure against a loss of the skill or health needed to generate income. These unhedgeable and uninsurable risks to the total income-earning capacity of a household, in combination with risk aversion, entail an increased incentive to save. In other words, even uncertainty about the security of assets can work to increase as well as to decrease savings, because in unfavorable contingencies assets contribute a large increment to prospective utility. This implication is corroborated even by everyday language, as when people speak of "saving for a rainy day." Second, the greater the amount saved, the smaller in general will be the ratio of the transactions costs of diversifying the porfolio or protecting it in other ways[5] to the value of the portfolio. In summary, uncertainty may reduce the expected apparent rate of return to saving, yet at the same time in the absence of universal futures and insurance markets it can increase the "need" or incentive for saving, that is, its prospective utility. Because, in addition, uncertainty may favor accumulation of enough wealth to

[5] In the case of households whose assets are acquired by illegal or unpopular means, or which become so large they generate notoriety, additional saving may if anything increase the possibility of prosecution, extortion, or adverse political decisions. Yet concealment and diversification can reduce these risks also, perhaps more easily as wealth grows larger.

make diversification or other forms of portfolio protection practical, it need not lower even the apparent rate of return.

The question of whether those uncertainties that imply the same behavior as positive time preference are quantitatively more significant than those that imply the same behavior as a negative time preference is an empirical question. Yet to the best of our knowledge it has not even been addressed. Nor are we aware of any persuasive empirical literature on whether other possible influences on saving behavior, such as irrationality or perceptual bias, combine with the two opposing effects of uncertainty to create an appearance of time preference that is on balance positive, or zero, or negative.

VI. Is Time Preference Positive?

The argument in prior sections helps us address this question now. We shall see in the next section that there are reasons for assuming that some households can have an indefinitely long or infinite time horizon, and that the conclusion we shall draw with this assumption holds a fortiori when there are shorter planning horizons. If $\eta = 0$, from equation (8) we have

$$\frac{v'(C_t)}{v'(C_0)} = \frac{1}{(1 + r)^t}. \tag{9}$$

If we then extend the time horizon to later and later t and apply (9) to the "last" period in the future, we obtain

$$\lim_{T \to \infty} \frac{v'(C_T)}{v'(C_0)} = 0. \tag{10}$$

This result implies that with the continuing positive interest rates that normally prevail and an η of zero, the household would lower C_0 to the desperation level, where $v'(C_0) = \infty$, if the C_T attainable under the budget constraint (7) is below the satiation level, at which $v'(C_T) = 0$. Inasmuch as we normally do not observe either people who cut consumption down to utterly abject levels in order to provide for the future or people who have because of past saving reached consumption levels at which their taste for goods is satiated, there appears to be a positive rate of time preference, that is, in our formulation an $\eta > 0$. (We discuss later a possible escape from this conclusion if there is *both* an infinite time horizon and what we call "drastically diminishing marginal utility.")

We have shown earlier that r is normally greater than η because capital is useful in production, and the present argument suggests that if r exceeded η by a very large amount that would also generate implausibly low levels of present consumption. This suggests that, at least in those societies where real interest rates are very high and

expected to remain so, η must exceed zero by more than a miniscule amount.[6]

Before turning to finite planning horizons, we should examine a familiar policy conclusion which assumes an infinite time horizon. This is the prescription of those who claim or imply that, at least when ecological systems or depletable resources are at issue, the only right and moral attitude is, in effect, a zero rate of time preference and an infinite time horizon—the interests even of distant generations should have the same weight as one's own interest in current-year consumption. One need not have a what-has-posterity-ever-done-for-me attitude or a shortsighted view to oppose this supposedly moral command. It implies, as a glance at (10) will remind us, that at every positive interest rate the household should reduce its consumption to the level at which the marginal utility of current consumption is infinite, in order to give all to the future. Of course, that applies to every future year, and to every future generation, until either income rises so high that the marginal utility of consumption (when consumption equals income) has fallen to zero, or alternatively there is so much capital around that the interest rate is no longer positive. We doubt that the proponents of a zero time preference and an infinite time horizon have understood the implications of their argument and conclude that their argument surely cannot be a proper basis for national policy.[7]

The existence of positive time preference also has immediate implications for the responsiveness of savings to changes in interest rates. We earlier defined η as zero for a household if, when it had equal endowments in each period and faced an interest rate of zero, it chose a stable level of consumption. A positive η in these conditions, or any η greater than the interest rate, implies a decreasing level of consumption through borrowing or dissaving, whereas an $r > \eta$ implies net saving. If $\eta = r$ at any level, there is a stable level of consumption. Starting from such a situation and looking at equation (8), we ask what would happen if the interest rate fell to zero. The analogy with the results in equations (9) and (10) immediately reminds us that this would give us

$$\frac{v'(C_t)}{v'(C_0)} = \frac{(1+\eta)^t}{1}. \tag{11}$$

As T gets indefinitely long, this would give

$$\lim_{T\to\infty} \frac{v'(C_T)}{v'(C_0)} = \infty. \tag{12}$$

[6] We are thankful to Peter Murrell for calling this point to our attention.

[7] For a related finding, see the Koopmans (1960) article discussed below.

We see that either C_0 increases until $v'(C_0) = 0$, or $v'(C_T)$ decreases until $v'(C_T) \to \infty$. That is, the ratio C_0/C_t will become huge: The family will dissave on a large scale, unless it somehow suffers a permanent plight so severe that $v'(C_t)$ goes rapidly boundless with a small reduction in C_t, or so near satiation of the capacity to enjoy goods that $v'(C_0) \to 0$ quickly. This shows that it is logically contradictory to hold that the representative household is myopic or for any other reason has positive time preference, and also to hold that savings do not respond to changes in the interest rate and would normally remain positive at a zero interest rate.[8] (In fact, saving would be sensitive to the rate of interest even if time preference were negative and, since the marginal utility of consumption is normally diminishing, also if it were zero in any environment in which endowments vary over time.) We think this implication may surprise many readers. It will be clear from the next two sections that this conclusion, like the others, also holds true with finite time horizons.

VII. Truncated Time Horizons

Most significantly, the case that there is positive time preference—that is, something beyond Böhm-Bawerk's "first cause"—gets stronger as the planning horizon gets shorter. The reason for this is utterly obvious once stated: A finite planning horizon of T years implies that whatever happens after T years is not only discounted but given a value of zero. Were there not usually markets in which one can trade with those whose time horizons extend farther into the future, an individual would place no value at all on an asset, such as a bond, that could be redeemed or sold only after the end of his planning horizon. Truncating a decision maker's time horizon can therefore only strengthen the evidence for a positive time preference, since it merely involves attributing a zero value, rather than a discounted but positive value, to any consumption after a certain date.

The fact that truncated time horizons lead so directly and simply to positive time preference makes one ask how common and how long truncated time horizons might be. Are observed savings rates, which as we have shown are far lower than a zero time preference and an infinite time horizon would imply, as low as they are simply because the behavior we actually observe is driven by time horizons on average about a lifetime or so in length?

The notion that time horizons might average about a lifetime in

[8] We have in mind a compensated saving schedule, along which household real income is constant, so that the effect of a change in the interest rate is a pure substitution effect.

length has some appeal at first glance. It certainly must be an overestimate for older people without heirs or other survivors about whom they are concerned. Probably it is about right for some others too, who for one reason or another may feel that the cares of one lifetime are enough. The number with concern for heirs is probably larger, but it appears that mainly this concern extends only a generation or two into the future. Even in dynastic families, such as that of the Capetian kings of France, we may wonder whether they thought ahead much beyond producing a male heir and preserving his birthright. Taking the different types of cases together, it might seem that truncated time horizons might average out to perhaps a lifetime, and that this is the main explanation of the positive η that we inferred from the small percentage of incomes that are saved.

This interpretation is nonetheless unsatisfactory, as we realize from our earlier reference to the fact that lifetime horizons imply no saving by societies, unless the age distribution for a time contains disproportionate numbers of people in the "savings" stages of life. There is substantial social saving, especially when we include education and human capital as we must, that is not explainable by growth, so the typical time horizon must be more than a generation in length.[9]

This consideration in turn means that there is no special need to know the rate of time preference within planning horizons of a lifetime or less in length. Though individuals with time horizons as short as this presumably exist, the evidence adduced above suggests that they are exceptional. It is conceivable that these exceptional individuals could have a zero or negative time preference over the course of their planning horizons. If so, there is a remarkable discontinuity, for we know from the definition of a finite planning horizon that nothing that occurs after it is even worth taking into account. The dominant fact is that the time preference for consumption within the planning horizon is infinite compared to anything that happens after it. Whether one is interpreting aggregate time series on saving and accumulation or analyzing social policy about them, this infinite time preference is the dominant fact, besides which the presence or absence of time preference within the life span pales to insignificance.

It is indeed fortunate that time preference within planning horizons of a lifetime or less in length is not the key issue, for (contrary to what we once believed) it is almost impossible to find out what time preference over the life cycle might be. Family size and optimal

[9] Growth due to technological advance in what would otherwise be a steady state also implies some net social saving, because young savers have higher lifetime incomes than have older people, so that their total saving exceeds the dissaving of retired people. However, White (1978) and Kotlikoff and Summers (in press) find that total saving far exceeds the amount that can be attributed to life-cycle saving in a growing economy.

periods for major investments in human capital change over the course of life, and though it is easy to incorporate these considerations into our theoretical framework, the changes in spending and savings levels they bring about are difficult to disentangle from any short-run time preference. Even the individual's capacity to enjoy consumption may vary with age, as is suggested by the cliché that one should not postpone high levels of consumption until one is "too old to enjoy it." In most households the wage rate also changes substantially with age, and this again affects the level of savings in each period. There are even problems arising because of changes in the value of time due to these changes in wage rates. The final outputs desired by consumers are characteristically produced through household production functions in which time and market goods are inputs. As Becker and Ghez (1975) have shown, changes in wages over time can affect savings levels in at least two significant ways. First, the higher the wage rate the greater the incentive to substitute purchased goods for time in the domestic production of final outputs, as may occur when a family increases its reliance on caterers, convenience foods, and restaurants when wage rates of family members are unusually high. Second, the anticipation of a different opportunity cost of time in the future can change the time profile of final consumption and savings, so that there is more consumption in periods when a lower cost of time makes the price of final output cheaper, as could occur when a family postpones the spending involved in distant tourism till retirement. Given the unknown and sometimes contrary effects of these and other influences, it is not immediately clear what levels of savings would support or refute the hypothesis that time preference is positive within time periods of a lifetime or less.

VIII. Bequests

Let us now return to the more representative case of a household with heirs for which it has some concern. What a couple during the course of life gives to its children and what it bequeaths to them has a capital value equal to the discounted future value of the consumption (and later bequests) it is expected to provide. In the minds of the givers this bequest must have a prospective value or utility at least equal to the utility forgone by not consuming it themselves. The extent of any willingness of the household to reduce the heads' current consumption to provide for their children's endowments is the central parameter revealing the presence or absence of the "myopia" discussed in debates about the determinants of saving.

A household with the faith that it will have a continuing line of descendants with the same regard for their children as it has for its own will act as if it had an indefinitely large or infinite time horizon, even if it gives such a long time horizon no explicit attention. Suppose that the household spends that amount on education, gifts, and bequests that strikes its preferred balance between its own desire to consume and its provision for the children's future, and leaves it to the children to allocate their resources in turn. When the problem is understood in this way it becomes clear that there is an extensive theoretical literature that is directly applicable, notably that of dynamic programming. One of the basic findings in this literature is that the optimal program for several periods is the same as that reached by optimizing each period in turn, including setting optimal end-of-period conditions each time for the beginning of the next period (see, e.g., Bellman and Kalaba 1965, chap. 2); the result is the same whichever way one looks at the problem. That this notion applies precisely to intergenerational resource allocation is shown by Barro (1974). It can accordingly be appropriate to view the choice of the size of the bequest as if the household were maximizing a utility function of the form (1) with an indefinitely large, or infinite, time horizon. The T represents not how far into the future people plan consumption, but the length of a dynamic program implied by a continuing series of intergenerational transfers.

Let us now suppose that the conditions that make it appropriate to assume an infinite time horizon do not hold, so that finite time horizons such as we used to develop equations (1)–(8) are appropriate. If the present heads of household are utterly unconcerned about any descendants beyond, say, their grandchildren, this lack of concern will manifest itself in the form of a smaller bequest than otherwise. Uncertainty about whether a subsequent generation will have offspring, or fail to make a bequest to its children, will also make for a smaller bequest. The effect of a lack of concern for more distant generations, or of uncertainty about whether the bequest will in fact serve the givers' purposes, is approximately the same as if the household applies a high rate of time preference η in maximizing utility (1) with a large or infinite T. We can see this by taking equation (1) and making $v(C_t)$ a declining function of time, for example, $v(C,t)$ with $v_t <$ 0. This is in effect equivalent to keeping (1) as it is and raising the value of η. This makes it clear in another way that less-than-infinite time horizons entail positive time preference. Thus the decision to let T in equation (1) be finite at some points, and to let it approach infinity in Section VI, was governed largely by analytical convenience and had no substantive effect on the conclusions.

IX. Drastically Diminishing Marginal Utility

With infinite time horizons and continued positive interest rates the returns to savings ultimately become so colossal that, if there were not positive time preference, current consumption would be drawn down to survival levels. There are, however, two phenomena in addition to positive time preference that conceivably could explain the high levels of consumption that actually occur. One is that such large increases in endowment income are expected that saving for still more future consumption is not worthwhile however much interest would accrue.[10] The other possibility is that the marginal utility from consumption in any period diminishes so drastically as consumption rises that even essentially unlimited interest income would not yield enough future utility to justify much saving. There is no reason for most households to assume increases in endowments so great that they could justify the low observed savings rates. It is, on the other hand, just possible that drastically diminishing marginal utility in combination with plausible expectations about increasing endowment incomes, perhaps due to technological advance, could explain the observed levels of saving.

Suppose, for example, that the household expects an exponentially rising real income due to technological advance, even if its saving and aggregate savings are zero. Equation (8) implies a chosen rate of growth of consumption that is independent of the horizon T; this growth will be exponential if marginal utility is a constant-elasticity function, that is, if utility has the form

$$v(C) = AC^{\alpha} + B \tag{13}$$

for $\alpha < 0$, $A < 0$, and $B > 0$, or, for the unit elastic case, the form

$$v(C) = A \ln C + B \tag{14}$$

with $A > 0$ (see White 1978). For Harrod-neutral technological advance, there is some fraction of income saved that is consistent with having consumption grow exponentially at the same rate ρ as income does. If so we can have $v'(C_T) \to 0$ as $T \to \infty$ because of this growth.

For example, suppose that utility is given by the Bernoulli function (14). Then the optimum consumption pattern implied by (8) is

$$C_t = \frac{(1 + \bar{r})^t}{(1 + \eta)^t} C_0 \tag{15}$$

for every t. If it should happen that $1 + \rho = (1 + \bar{r})/(1 + \eta)$, or, approximately,

[10] We are grateful to Christopher Clague for calling our attention to this possibility.

$$\rho = \bar{r} - \eta, \tag{16}$$

consumption will grow at the same rate as income and will remain a constant fraction of income, as is actually observed. This pattern can be sustained indefinitely and can plausibly be reconciled with a constant rate of interest (see Solow 1970, chap. 5).

If the rate of time preference were zero, equation (15) would imply that consumption rises indefinitely at the rate $\bar{r}$. When $\bar{r} > \rho$, as has been true of the United States, this high rate of growth of consumption can be maintained for large T only by pushing C_0 down toward zero, pushing the fraction of income saved toward unity.

This set of results and considerations makes a compelling case for the existence of a positive rate of time preference, unless one supposes that the utility function $v(C)$ is inelastic and that unlimited exponential growth of income is expected even with zero saving. In the past, output per unit of input has grown at about 1 percent a year, and the U.S. birthrate has converged on zero population growth. If the real rate of interest of about 5 percent (which has prevailed during most of the past several decades) is what households expected to continue into the indefinite future, these figures can be shown to reconcile with zero time preference only if the elasticity of the utility function (13) is minus 5 corresponding to $\alpha = -4$. This function has what we shall label "drastically diminishing marginal utility." We do not argue that utility functions of this character are the norm but know of no data that rule them out and think it is not without interest that (when exogenous increases in endowments are expected and time horizons are also infinite) the facts of saving force us to choose between positive time preference and drastically diminishing marginal utility of income.

X. The Relationship to Prior Findings

Though many of the prior findings on time preference contradict one another, and there has been no generally accepted analysis of the matter, many of these writings nonetheless have considerable value. With the aid of the findings here we are in a better position to understand or appreciate some prior contributions.

One of the best-known contributions is that of Strotz (1956). The extreme simplicity of Strotz's insight and its relationship to our argument can be made immediately apparent. Suppose that an individual has a positive time preference, in that he values consumption in the future less highly than consumption now, but it is only the distance into the future, rather than the specific date, that determines the weight given to the utility. Suppose for the sake of a simple example that the marginal utility of consumption in any one period is the

same at all levels of consumption, however high or low, so that we can ignore the diminishing marginal utility of consumption and focus exclusively on Strotz's insight. Let us suppose that our individual must obtain all consumption out of an initial endowment of 25 and that his time preferences are such that he chooses initially to consume 20 units, or four-fifths of his stock, in the present period, three units in the second period, and two units in the third. When the first period is over, however, the individual will abandon his initial plan and consume four-fifths of his remaining endowment of five, or four, since by stipulation he weighs the utility in the present so highly he consumes four-fifths of his stock, and period 2 is now the present. This is of course inconsistent with the initial plan for consumption in the second and third periods; the initial plan has not been followed, even though all of the individual's expectations about his situations and tastes have proved correct. The point is that no individual with the pattern of time preference described will voluntarily adhere to the plans he makes in advance; it is impossible for him to maintain his pattern of time preference and also adhere to the intertemporal plans he makes at the beginning. It becomes obvious on reflection that the only patterns of positive time preference that individuals can have and still make consistent or rational intertemporal choices are those which involve a constant exponential rate of discounting: a constant η such as we have all along assumed. If an individual plans to save the same percentage of his period 2 stock as of his period 1 stock, he will in fact carry through with that plan if his tastes do not change. Of course, no positive time preference at all ($\eta = 0$) is consistent with rational behavior; it would have led our hypothetical individual to consume a third of his stock in each of the periods and to have adhered to that plan. Thus we see that it is not consistent to assume both rational behavior and positive time preference without also assuming that η is an exponent of constant value.

Unfortunately, Strotz goes beyond the insight just described to define "prudent" or "thrifty" behavior as any behavior which meets the standard of intertemporal consistency: An individual is arbitrarily defined to be "thrifty" if he discounts the future at a constant exponential rate, however high that rate may be. In keeping with his preoccupation with perceptual error, inconsistent choices, and the psychology of intertemporal choices, Strotz concludes that "consumer sovereignty has no meaning in the context of the dynamic decision-making problems" (1956, p. 179).

Though Strotz's disturbing conclusion is consistent with the rest of his argument, we can now readily see that it is not in fact justified. If a positive time preference, which he calls "myopia," were solely the result of irrationality and perceptual errors, then consumer sovereignty would indeed have little or no meaning for intertemporal

choice. But as we have seen, the absence of a positive time preference implies extreme behavior, or else drastically diminishing marginal utility. Short planning horizons can occur and need not imply any irrationality, and the very act of ignoring any return beyond some limited planning horizon as we know entails a positive time preference. Thus, though the intransitive and irrationally impatient behavior Strotz illuminated can of course also occur, his nihilistic general conclusion about the inapplicability of consumer sovereignty to intertemporal choices is incorrect.

Our analysis is in closer accord with writings of Koopmans, Diamond, and Williamson (1964) and Hirshleifer (1970). These writers also conclude that there is positive time preference, which they define somewhat differently. They do not, however, offer any "empirical test" of the sort we used, nor go into shorter time horizons and drastically diminishing marginal utility. Koopmans et al. (1964) prove that various sets of formal postulates entail the existence of positive time preference. The formal postulates are naturally chosen in part because of their tractability, and it is not always immediately clear whether these postulates are very restrictive or succeed in capturing the essence of the matter. It is not immediately clear, for example, whether Diamond's (1965) assumption that each period's utility $v(\cdot)$ is bounded is limiting, or even material to the proof, or whether it matters that these authors exclude the possibility that $v'(C_0) = \infty$ or that $v'(C_t) = 0$, which we allow. The fact that more work needs to be done before the generality and applicability of the Koopmans (1960) and Diamond (1965) proofs can be definitely established is perhaps best demonstrated by quoting Koopmans's briefest summary of the reason for positive time preference in his article " . . if there is in all circumstances a preference for postponing satisfaction—or even neutrality toward timing—then there is not enough room in the set of real numbers to accommodate and label numerically all the different satisfaction levels that may occur in relation to consumption levels for an infinite future" (1960, p. 288). Hirshleifer (1970, p. 96), using a line of reasoning much closer to ours, concludes in an important if cryptic passage that zero time preference would imply infinite deferral of all consumption. That is closer to the mark but overlooks some of the possibilities. In general, if the argument in our paper is correct, it suggests that the contributions we have just cited deserve more attention and perhaps further extension as well.

More recently, Stigler and Becker (1977) have put forth conclusions that directly contradict those of Strotz (1956), Koopmans et al. (1964), and Hirshleifer (1970). Stigler and Becker argue that many types of behavior that economists have traditionally characterized as special patterns of taste are in fact explicable in terms of different prices and incomes. One of the ad hoc assumptions about tastes to which Stigler

and Becker especially object is the assumption, going back as we know to Böhm-Bawerk, that people "'systematically undervalue . . . future wants.' The taste for consumption in say 1984 is alleged to continue to shift upward as 1984 gets closer to the present. In spite of the importance frequently attached to time preference, we do not know of any significant behavior that has been illuminated by this assumption. Indeed, given additional space, we would argue that the assumption of time preference impedes the explanation of life cycle variations in the allocation of resources, the secular growth in real incomes, and other phenomena" (Stigler and Becker 1977, p. 89).

It might seem that the quoted passage by Stigler and Becker contradicts our own findings no less than those of the authors we have cited, since we also claim to have demonstrated the existence of positive time preference. This is not in fact the case; one could consistently accept Stigler and Becker's conclusion and our own as well. We have defined positive time preference in a slightly broader and more readily testable fashion than Stigler and Becker have, and we neither assumed nor ruled out irrationality or perceptual error. Accordingly, it is, as we argued, an empirical question whether there is time preference, and the revealed preference approach made it possible to assess the evidence. It would of course be possible to discuss the question of why bequests are as small as they are, or why some people have only finite planning horizons, with different terminology, though excluding the phrase "time preference" and synonyms for it would in our judgment involve circumlocution.[11]

XI. Conclusions

Our principal findings are the following:

1. A finite time horizon implies that only consumption inside the horizon has value and signifies an absolute unwillingness to transfer goods to children or other descendants who could benefit from them after a specified future date. Thus a finite planning horizon is both theoretically and practically equivalent to a positive rate of pure time preference. Planning horizons that do not reach beyond the lifetimes of those who make the plans can exist, and (though this implies a remarkable discontinuity) it is logically possible that within such lifetime horizons time preference could be zero or even negative. But this possibility can have little significance in comparison with the high rate of time preference implied by the short horizon, and in any event it does not seem feasible to disentangle time preference during

[11] Stigler and Becker's footnote assertion that "a consistent application of the assumption of stable preferences implies . . . the absence of time preference" (1977, p. 78, n. 4) is either mistaken or dependent upon definitions of "stable preferences" or "time preference" that appear to be untypical.

the course of a lifetime from other changes over the life cycle. Time horizons which imply no bequests for descendants also cannot be the norm, since they would imply lower savings than are observed and no social saving whatever in a society in a stationary state. We have also shown that finite time horizons and positive time preference are functionally equivalent, so the most interesting question is whether there is positive time preference within an infinite time horizon. If there is, there is no need to go into the length of planning horizons, and the analytically convenient assumption of an infinite time horizon can be used without significant loss of generality.

2. An infinite time horizon and a zero rate of time preference suggest that (if the foreseeable rate of economic growth is less than the expected long-term rate of interest) current consumption would be no higher than the subsistence level. Over an indefinitely long time span continued positive interest rates imply such colossal gains in future consumption from reductions in current consumption that all income beyond that needed for survival would be saved.

3. There is, however, the alternative possibility that even vastly higher consumption is expected to bring so little increase in utility that households are unwilling to forgo consumption now, even though this can generate essentially limitless increases in consumption later. The possibility that there will be absolute satiation at these future consumption levels can be rejected on the classical grounds that wants are limitless, but there remains the intriguing possibility that there is what we here labeled "drastically diminishing marginal utility": Even boundless increases in future consumption would then bring such modest increases in utility that the low levels of saving we observe could be rational even with a zero rate of time preference. The observed levels of saving imply positive time preference, or drastically diminishing marginal utility, or both of these interesting phenomena. In an age that has seen secularly rising incomes, it might seem that drastically diminishing marginal utility is a type of positive time preference, since an unwillingness to bequeath much to descendants, on the grounds that they will be better off than we are and will not get much satisfaction from still higher consumption, provides a good rationale for preferring current consumption and leads to the same observed savings levels. This is not in fact correct. If there were a sufficient technological regression or other losses in endowments, yet a positive interest rate, zero time preference would imply that current consumption would be reduced to subsistence levels, but drastically diminishing marginal utility would not. Similarly, if for shifts of goods within their own lifetimes people behave as if they had Bernoulli utility functions, intertemporal consistency would require the use of positive time preference to reconcile this behavior with their gifts and bequests. Also, whether drastically diminishing utility is

present or not, finite planning horizons entail positive time preference. Drastically diminishing marginal utility also may be needed to explain other phenomena, such as a willingness to pay a high price to insure against even modest losses or to reduce fluctuations in consumption levels, and savings levels in particular cases can be traced to drastically diminishing marginal utility according as its other implications are present or absent.

4. Uncertainty about the future has double-edged effects. Of course, if the typical household thinks that there is a greater likelihood that part of its bequest will merely be destroyed in a future catastrophe than that the full bequest adds to the probability that descendants will survive well in that catastrophe, the expected rate of return to saving is thereby lowered in their minds. That expectation, instead of time preference, could in principle account for high consumption and low saving. However, each type of uncertainty we can think of offers possible reasons for saving more as well as reasons for saving less. As in the other cases, it is hard to see a simple way to distinguish it empirically from pure time preference. Thus for purposes of this inquiry we have made the simplest assumption: that uncertainty about whether any marginal savings will be lost is on average about offset by the incentives for additional saving that uncertainty about the future creates. To make a better specification possible in any further studies, we have made clear above that the prevalence of futures and insurance markets, the degree of risk aversion, the degree of confidence in predictions about the future, and the costs of diversification or protection of portfolios each has distinct effects on savings rates, and these might eventually be estimated.

In summary, the case for positive time preference is absolutely compelling, except in the interesting case where there is drastically diminishing marginal utility. Utility functions of this sort have interesting implications of their own which need to be tested separately. Time preference is accordingly by no means a subject that deserves only casual references to alleged perceptual errors or erroneous inferences from the existence of positive interest rates. It is a phenomenon of great practical and theoretical importance that is fortunately quite amenable to analysis.

References

Bailey, Martin J., and Olson, Mancur. "Pure Time Preference, Revealed Marginal Utility, and Friedman-Savage Gambles." Working Paper no. 77-2, Univ. Maryland, 1977.

Bailey, Martin J.; Olson, Mancur; and Wonnacott, Paul. "The Marginal Utility of Income Does Not Increase: Borrowing, Lending, and Friedman-Savage Gambles." *A.E.R.* 70 (June 1980): 372–79.

Barro, Robert J. "Are Government Bonds Net Wealth?" *J.P.E.* 82, no. 6 (November/December 1974): 1095–1117.

Becker, Gary S., and Ghez, Gilbert R. *The Allocation of Time and Goods over the Life Cycle.* New York: Columbia Univ. Press (for Nat. Bur. Econ. Res.), 1975.

Bellman, Richard E., and Kalaba, Robert. *Dynamic Programming and Modern Control Theory.* New York. Academic Press, 1965.

Bohm-Bawerk, Eugen von. *Capital and Interest.* Vol. 2, *Positive Theory of Capital.* South Holland, Ill.: Libertarian Press, 1959.

Debreu, Gerard. "Topological Methods in Cardinal Utility Theory." In *Mathematical Methods in the Social Sciences*, edited by Kenneth J. Arrow, Samuel Karlin, and Patrick Suppes. Stanford, Calif.: Stanford Univ. Press, 1959.

Diamond, Peter A. "The Evaluation of Infinite Utility Streams." *Econometrica* 33 (January 1965): 170–77.

Dorfman, Robert. "A Graphical Exposition of Bohm-Bawerk's Interest Theory." *Rev. Econ. Studies* 26 (February 1959). 153–58. (*a*)

——— "Waiting and the Period of Production." *Q.J.E.* 73 (August 1959). 351–72. (*b*)

Feldstein, Martin S. "The Social Time Preference Discount Rate in Cost Benefit Analysis." *Econ. J.* 74 (June 1964): 360–79.

Frisch, Ragnar. "A Complete Scheme for Computing All Direct and Cross-Demand Elasticities in a Model with Many Sectors." *Econometrica* 27 (April 1959): 177–96.

Hirshleifer, Jack. *Investment, Interest, and Capital.* Englewood Cliffs, N.J.: Prentice-Hall, 1970.

Jevons, W. Stanley. *The Theory of Political Economy.* 5th ed. New York. Kelley, 1965.

Koopmans, Tjalling C. "Stationary Ordinal Utility and Impatience." *Econometrica* 28 (April 1960): 287–309.

Koopmans, Tjalling C.; Diamond, Peter A., and Williamson, Richard E. "Stationary Utility and Time Perspective." *Econometrica* 32 (January–April 1964): 82–100.

Kotlikoff, Laurence J., and Summers, Lawrence H. "The Role of Intergenerational Transfers in Aggregate Capital Accumulation." *J.P.E.* (in press).

Marglin, Stephen A. "The Social Rate of Discount and the Optimal Rate of Investment." *Q.J.E.* 77 (February 1963). 95–111.

Pigou, Arthur C. *The Economics of Welfare.* 4th ed. London. Macmillan, 1960.

Ramsey, Frank P. "A Mathematical Theory of Saving." *Econ. J.* 38 (December 1928). 543–59.

Samuelson, Paul A. "A Note on Measurement of Utility." *Rev. Econ. Studies* 4 (February 1937): 155–61.

Solow, Robert M. *Growth Theory: An Exposition.* New York. Oxford Univ. Press, 1970.

Stigler, George J., and Becker, Gary S. "De Gustibus Non Est Disputandum." *A.E.R.* 67 (March 1977). 76–90.

Strotz, Robert H. "Myopia and Inconsistency in Dynamic Utility Maximization." *Rev. Econ. Studies* 23, no. 3 (1956): 165–80.

White, Betsy Buttrill. "Empirical Tests of the Life Cycle Hypothesis." *A.E.R.* 68 (September 1978): 547–60.

Temporary Income Taxes and Consumer Spending

Alan S. Blinder

Princeton University and National Bureau of Economic Research

Both economic theory and casual empirical observation of the U.S. economy suggest that spending propensities from temporary tax changes are smaller than those from permanent ones, but neither provides much guidance about the magnitude of this difference. This paper offers new empirical estimates of this difference and finds it to be quite substantial. The analysis is based on an amendment of the standard distributed lag version of the permanent income hypothesis that distinguishes temporary taxes from other income on the grounds that the former are "more transitory." This amendment, which is broadly consistent with rational expectations, leads to a nonlinear consumption function. Though the standard error is unavoidably large, the point estimate suggests that a temporary tax change is treated as a 50-50 blend of a normal income tax change and a pure windfall. Over a 1-year planning horizon, a temporary tax change is estimated to have only a little more than half the impact of a permanent tax change of equal magnitude, and a rebate is estimated to have only about 38 percent of the impact.

In 1968, faced with a classic case of demand inflation, Congress enacted a temporary increase in personal income tax payments to

This paper has benefited from seminar presentations at Princeton, Hebrew University of Jerusalem, Tel-Aviv University, and the University of Pennsylvania, and from helpful suggestions from Walter Dolde, Marjorie Flavin, Roger Gordon, Levis Kochin, Robert Solow, James Trussell, and a referee of this *Journal*. Most of the laborious computations were done expertly by Suzanne Heller, additional research assistance was provided by David Card, Robin Lindsey, and William Newton. Stephen Goldfeld and Richard Quandt generously guided this novice through the intricacies of their nonlinear optimization routines. Financial support from the National Science Foundation and from the Institute for Advanced Studies in Jerusalem, where part of this work was done, is gratefully acknowledged. This research is part of the National Bureau of Economic Research's program in economic fluctuations, but any opinions expressed herein are those of the author, not of the NBER.

[*Journal of Political Economy*, 1981, vol. 89, no. 1]
© 1981 by The University of Chicago 0022-3808/81/8901-0003$01.50

curb aggregate demand. In 1975, near the trough of our worst postwar recession, Congress enacted a tax rebate and other temporary decreases in taxes and increases in transfer payments designed to stimulate aggregate demand. Questions have been raised about the effectiveness of both measures.

The questions have both theoretical and empirical roots. On theoretical grounds, the permanent income–life cycle hypothesis seems to argue that temporary income tax changes should have little effect on consumer spending in principle. On empirical grounds, data on consumer behavior seem to suggest that the impacts of the two temporary taxes on spending were also small in practice. In 1968, the savings rate fell from 7.5 percent in the quarter immediately preceding the tax surcharge (1968:2) to only 5.6 percent in the first quarter of the surtax, suggesting that consumers kept spending despite the tax. In 1975, the savings rate ballooned from 6.4 percent just prior to the rebate to a stunning 9.7 percent in the quarter of the rebate (1975:2), suggesting that little of the rebate was spent.

The purpose of this paper is to study these two temporary tax changes in some detail. Precisely what prediction does economic theory make about the relative effectiveness of temporary versus permanent tax changes? And what conclusions can be reached from U.S. time-series data? The fact that the Carter administration asked for (but did not get) a repeat performance of the rebate in 1977 suggests that there is more than academic interest in the answers to these questions.

Section I outlines the theoretical issues, beginning with an idealized life-cycle model and proceeding to introduce some important "real world" considerations. Since the discussion shows quite clearly that the issue is an empirical one, Section II reviews previous empirical work on the subject very briefly. Section III explains the underlying basis of the empirical model of this paper, relating it to recent literature on rational expectations and the permanent income hypothesis (PIH), and then Sections IV and V show how this basic conceptual framework was converted into an operational empirical model. The estimates are presented and analyzed in Section VI, and Section VII summarizes the main conclusions.

I. The Implications of Theory: Pure and Impure

1. *The Pure Permanent Income–Life Cycle Theory*

As Eisner (1969) pointed out some time ago, the PIH casts doubt on the effectiveness of income tax changes that are labeled as temporary because such measures have only minor effects on permanent income.

To develop a theoretical benchmark for the marginal propensity to consume (MPC) that the PIH suggests might apply to a temporary tax, consider a rarefied world in which consumers with exogenous earnings streams select consumption paths to maximize lifetime utility. If capital markets are perfect, only the discounted present values of the earnings streams matter, so suppose all households earn a constant income y per year. Assume further that households differ only in age, a; that the real rate of interest is zero; and that the subjective rate of time discounting is also zero.[1] The question is, If income taxes are raised by z per capita for the period from $t = 0$ to $t = t_1$, how much less will consumers spend over this interval?

In answering this question, there are three population groups to keep track of. People who are "alive" (in the economic sense) at $t = 0$ and who live past the expiration of the tax suffer an income loss of $t_1 z$ over the period. If T denotes the length of life, then each such person of age a consumes a fraction $t_1/(T - a)$ of this loss during the years in which the temporary tax is in effect. Thus the change in consumption *per capita* is $\Delta C_1 = zt_1^2/(T - a)$.

Old people who are alive at $t = 0$, but die before $t = t_1$, lose only $(T - a)z$ in income. However, since they have MPCs of unity during the surtax period as a whole, their change in consumption per capita is $\Delta C_2 = (T - a)z$.

Finally, we must worry about people who are born between $t = 0$ and $t = t_1$. If a, a negative number between 0 and $-t_1$, denotes the age of such a person, and he lives for $t_1 + a < t_1$ years during the period $0 \leq t \leq t_1$, his income loss is $(t_1 + a)z$. Since he spends a fraction $(t_1 + a)/T$ of this income during the period $0 \leq t \leq t_1$, the change in his consumption is $\Delta C_3 = [z(t_1 + a)^2]/T$.

To derive the aggregate change in per capita consumption, weight these groups by the age distribution, considering the ages $-t_1 \leq a \leq T$. In the simplest case of a uniform age distribution, $f(a) = 1/T$, the change is

$$\Delta C = \int_0^{T-t_1} \frac{\Delta C_1}{T} da + \int_{T-t_1}^{T} \frac{\Delta C_2}{T} da + \int_{-t_1}^{0} \frac{\Delta C_3}{T} da.$$

Working out the integrals and dividing by the total income that is taxed away during the period ($t_1 z$ per capita), we obtain

$$\text{MPC} = \frac{t_1}{T}(\log T - \log t_1) + \left[1 - \frac{T}{2t_1} + \frac{(T - t_1)^2}{2Tt_1}\right] + \frac{t_1^2}{3T^2},$$

where the three terms show the contributions of the three different population groups.

[1] This is essentially the model introduced by Modigliani and Brumberg (1954).

To take a concrete example, suppose the typical lifetime of a household head as a household head is $T = 50$ years. Then, according to this formula, the MPC for a 1-year tax ($t_1 = 1$) is .09, while that for a 2-year tax ($t_1 = 2$) is .15.[2] The MPC = .09 for a 1-year temporary tax is only a rough benchmark representing the pure PIH, and there are a number of reasons why the theory probably systematically understates the responses of consumers to temporary taxes (see below), so we should not take this number too seriously. Still, there are two lessons worth drawing from this simple exercise—lessons that have often been forgotten in the temporary-tax debate.

a) Income gains and losses from temporary taxes will eventually be spent just like any other increment or decrement to lifetime resources: if less is spent at first (because $t_1 < T$), then more will be spent later. Thus if we want to inquire about the "effectiveness" of temporary taxes, we must specify a time horizon. Over a long enough run, they must be just as "effective" as permanent ones.

b) The so-called zero effect view—that consumers ignore the surtax and consume as if it never happened—does not represent the PIH at all. Instead, that theory says that consumers should spend precisely what they would on receipt of a windfall gain (or loss) of t_1z. In the illustrative calculation, this turns out to be the "9 percent effect" view.

2. *Caveats and Imperfections*

There are several reasons why surtaxes may affect spending more strongly than indicated by pure theory. First, tax-induced income changes that are not consumed must be saved. If windfall gains are used to purchase durable goods, consumer spending may rise much more strongly than consumption; the converse may happen when there are windfall losses. The magnitude of the marginal propensity to spend windfalls on durable goods is, of course, an empirical question.[3]

Second, some households may be subject to liquidity constraints that are usually ignored by the PIH. If we stay within the certainty context, these constrained households will react strongly to even temporary income changes.[4] Thus the aggregate MPC for a temporary tax is a weighted average of the low MPCs of unconstrained households and the high MPCs of constrained ones. Again, the importance of this phenomenon is an empirical question.

[2] By way of comparison, as $t_1 \to T$, the MPC $\to 5/6$

[3] See e.g., Darby (1972)

[4] See Blinder (1976) and Dolde (1978). For a look at one particular type of uncertainty, see Foley and Hellwig (1975), which shows that this result may not carry through to the uncertainty case.

Third, as Okun (1971) pointed out, consumer behavior depends on what people believe rather than on what the government announces. If consumers disbelieve the government when it tells them that a tax hike is only temporary, then the spending response will be greater than that suggested by a naive application of the PIH.[5] Since the perceived duration of the surtax, not the declared duration, is relevant from the standpoint of the PIH, this too raises an empirical issue.

Finally, we must recognize the possibility that households may not do the kind of rational long-term planning envisioned by Modigliani and Brumberg (1954) and Friedman (1957) or, what amounts to the same thing, have very high subjective discount rates. If they are very shortsighted, then temporary fluctuations in disposable income may have substantial effects on spending.

II. Previous Empirical Work

Okun's (1971) study opened the empirical debate on this issue. Using the consumption equations of four econometric models, he compared the "full effect" view that the 1968 surtax was just as effective as a permanent tax increase to the "zero effect" view that consumers totally ignored the surtax. While he concluded that the full effect view fit the data better, an intermediate "50 percent effect" view actually does better than either extreme.[6] Springer (1975) criticized Okun's econometric procedures and then performed a similar experiment with a consumption function based on the PIH. He concluded that the zero effect view performed better.

Juster (1977), using a series of savings equations based on the Houthakker-Taylor (1966) model, reached conclusions about the 1975 rebate similar to those of Okun for the 1968 surtax. But Modigliani and Steindel (1977) found that the rebate had very little impact over a horizon of 1 or 2 quarters. Their modified version of the life-cycle model implied, however, a virtually full effect view over a 6-quarter horizon. Modigliani and Steindel assumed that the nonrebate portions of the 1975 tax cuts were treated like permanent taxes and handled the 1968 surtax with dummy variables.

The existing empirical literature thus offers little consensus. The issue seems quite open.

III. The Distributed Lag Model of Consumption

The basic vehicle for investigating the effectiveness of temporary income taxes in this paper is the distributed lag version of the PIH.

[5] This point has rather less cogency with respect to tax cuts; but here too consumers may believe them to be more permanent than the government announces.

[6] On this, see Blinder and Solow (1974, pp. 107–9)

While this is the standard way of implementing the PIH empirically, the recent literature on rational expectations has seemed to raise doubts about its validity.[7] This section shows how the PIH and rational expectations together lead to an estimating equation very much like the one I use in this paper.

As Muth (1960) pointed out, the PIH is basically a *forward-looking* model of consumer behavior. It states that consumers, in deciding on their current spending, weigh their current asset holdings, their current income from labor, and their expected future income from labor. Specifically, permanent income at time t is defined as

$$Y_t^p = A_t + \sum_{s=0}^{\infty} \frac{{}_tY_{t+s}}{(1+r)^s}, \tag{1}$$

where A_t is the stock of real assets at the beginning of period t, Y_t is noninterest income in period t, and ${}_tY_{t+s}$ is the mathematical (i.e., rational) expectation of Y_{t+s} that is formed at time t. (By convention, ${}_tY_t = Y_t$.) However, if the stochastic process generating income can be described by a time-series model such as

$$Y_t = a_1Y_{t-1} + a_2Y_{t-2} + \ldots + a_{n+1}Y_{t-n-1} + \epsilon_t, \tag{2}$$

where ϵ_t is a white noise error term, then the resulting empirical model of consumption will be *backward looking*. For example, if the theoretical consumption function is

$$C_t = \delta + kY_t^p + u_t, \tag{3}$$

then the empirical consumption function will be

$$C_t = \delta + kA_t + b_0Y_t + b_1Y_{t-1} + \ldots + b_nY_{t-n} + u_t. \tag{4}$$

To see this, it is only necessary to note that the (rational) expectation of income in period $t + s$ can, in view of (2), be based only on the information set $\{Y_t, Y_{t-1}, Y_{t-2}, \ldots\}$. Thus, for example,

$$\begin{aligned} {}_tY_{t+1} &= a_1Y_t + a_2Y_{t-1} + a_3Y_{t-2} + \ldots + a_{n+1}Y_{t-n}, \\ {}_tY_{t+2} &= a_1{}_tY_{t+1} + a_2Y_t + a_3Y_{t-1} + \ldots + a_{n+1}Y_{t-n+1}, \\ &= (a_1^2 + a_2)Y_t + (a_1a_2 + a_3)Y_{t-1} + \ldots, \end{aligned}$$

and so on. Substituting all such expressions into the definition (1) and then into the consumption function (3), it is clear that (4) is derived. As pointed out by Sargent (1978) and others, the coefficients b_i in equation (4) will be complicated functions of the coefficients a_i in (2).

Suppose then that, as suggested by Dolde (1976), we can distinguish among two or more sources of income whose generating functions (2) may differ. The PIH in conjunction with rational expectations then

[7] See Lucas (1976) and esp. Hall (1978).

implies that the b's in (4) should follow a different pattern for each income source. A simple example will illustrate this point and also give us some feeling for possible magnitudes. Consider several income sources, each of which is generated by a first-order autoregressive:

$$Y_{it} = \rho_i Y_{i,t-1} + \epsilon_{it}. \tag{5}$$

Working out the expectations and plugging into (1) gives us a simple expression for the permanent income attributable to each source:

$$Y_{it}^{p} = \frac{1 + r}{1 + r - \rho_i} Y_{it}, \qquad \text{if } \rho < 1 + r. \tag{6}$$

Notice that, despite the long time horizon contemplated by the PIH, consumption depends only on current income. In general, consumption will depend on past income only up to lag n, where $n + 1$ is the longest lag considered in equation (2).

Now compare two income sources, one of which is entirely permanent ($\rho = 1$) and the other of which is entirely transitory ($\rho = 0$). A \$1.00 increase in the permanent component will, according to (6), raise permanent income by $\$(1 + r)/r$ and thus raise consumption by $\$k(1 + r)/r$. This may imply a very large immediate spending response.[8] By contrast, a \$1.00 fluctuation in the purely transitory component will, again according to (6), raise permanent income by only \$1.00 and thus raise consumption by only $\$k$. The lesson, of course, generalizes and applies far beyond the confines of first-order autoregressives: *income sources deemed to be more permanent will elicit prompter spending responses than income sources deemed to be more temporary.* The application of this principle to permanent versus temporary changes in taxes is apparent and immediate and was elucidated clearly by Lucas (1976). It is the basic notion underlying the empirical model to be developed in the next section.

However, lest confusion arise, I should stress that there is no sense in which the rationality of expectations is either assumed or imposed in the consumption functions estimated here. My point is only that the distributed lag formulation of the PIH is consistent with rational expectations. As Sargent (1978) has emphasized, rational expectations delivers a set of restrictions across equations (2) and (4) that can be imposed in estimating the two jointly. I have made no attempt to impose these restrictions here because my interest was in getting the best possible consumption-function estimates, not in testing rationality. Furthermore, it is well known that quite different models

[8] Suppose the rate of subjective time discounting is equal to the rate of interest, so that a constant consumption stream is optimal, and that B is the lifetime propensity to consume (i.e., $1 - B$ is the propensity to bequeath). Then k will be $Br/(1 + r)$, so that $k(1 + r)/r$ will be B, which is close to unity.

of consumption behavior (e.g., habit persistence) can lead to an estimating equation very much like (4). It is not my purpose to discriminate among alternative ways of arriving at (4).

IV. Derivation of an Estimating Equation

The preceding discussion makes it clear that different distributed lag coefficients might be associated with different sources of income. While the actual empirical analysis considered four types of income, the model is most readily explained if I suppose there are only two: income (positive or negative) attributable to temporary tax measures, which I denote as S_t ("special income"); and all other disposable income ("regular income"), which I denote as R_t. The R_t should not be confused with permanent income, since it has both permanent and transitory components. The basic idea underlying the estimating equation is that S_t is identifiably "less permanent" than R_t.

Suppose consumption responds to R_t according to a set of distributed lag weights: $w_j = \partial C_t/\partial R_{t-j}$, $j = 0, 1, \ldots, n$. Since the w_j depend on the stochastic process generating R_t, it is worth reporting that the deviations of R_t from a logarithmic time trend are well described by the following second-order autoregressive:[9]

$$y_t = \underset{(.09)}{1.28y_{t-1}} - \underset{(.09)}{.35y_{t-2}}, \qquad R^2 = .91, \text{ D-W} = 1.86.$$

When income follows a second-order autoregressive, permanent income as defined in (1) is:

$$Y_t^p = A_t + K_t + \frac{(1+r)^2}{(1+r)(1+r-a_1) - a_2} y_t + \frac{a_2(1+r)}{(1+r)(1+r-a_1) - a_2} y_{t-1},$$

if $a_1 + a_2/(1+r) < 1 + r$, where K_t is the present value of the trend component of labor income and y_t and y_{t-1} are current and lagged deviations from trend. Given the estimates of a_1 and a_2 above, and for $r = .0074$ (a 3 percent annual real interest rate), the implied coefficients are $Y_t^p = A_t + K_t + 13.5y_t - 4.7y_{t-1}$. This leads us to expect a very large value of w_0, followed by swiftly declining w's—possibly even turning negative. The empirical results bear this out.

As Lucas (1976) has argued, income changes that are clearly "more temporary" than regular income should get different spending

[9] Standard errors are in parentheses. Longer autoregressives, however, give slightly better fits. E.g., if $t - 11$ is the longest lag allowed to enter the regression, significant coefficients are obtained at lags 1, 2, 3, 4, and 11. An F-test for the zero restrictions implied by the second-order model, however, yields an F-ratio of only 1.77, which is well below the critical 5 percent point of the χ_9^2 distribution.

coefficients. To develop a model of the distributed lag response of C_t to S_t, first break down S_t into its components:

$$S_t = S_t^1 + S_t^2 + \ldots + S_t^m, \tag{7}$$

where S_t^i indicates the income gain or loss in quarter t from the ith temporary tax. (In the empirical work, $m = 3$.) It will help clarify the treatment of the S_t^i if I define a hypothetical set of lag coefficients β_j as the effect on C_t of a \$1.00 pure windfall gain received in quarter $t - j$.

The treatment of S_t^i depends on whether or not the ith temporary tax is still in effect. If it is, I assume that S_t^i is treated as a weighted average of regular and windfall income, so that it gets the distributed lag weights

$$\gamma_j = \frac{\partial C_t}{\partial S_{t-j}^i} = \lambda w_j + (1 - \lambda)\beta_j, \qquad \begin{array}{l} j = 0, 1, \ldots, n \\ 0 \leq \lambda \leq 1. \end{array} \tag{8}$$

If the temporary tax is no longer on the books, I assume that consumers look upon S_{t-j}^i in retrospect as if it had been a pure windfall and so apply the distributed lag coefficients β_j. By introducing a dummy variable defined as:

$D_t^i = 1$ if the ith temporary tax remains in force in quarter t,
$= 0$ otherwise,

it is possible to combine these two hypotheses into a single expression:

$$\gamma_j^i(t) = D_t^i[\lambda w_j + (1 - \lambda)\beta_j] + (1 - D_t^i)\beta_j, \tag{9}$$

where the notation now indicates that the γ weights depend both on calendar time and on the specific tax under consideration (because of the dummy variable).

An interesting point arises here. Standard pre-rational-expectations approaches to consumption-function estimation would suggest that $\Sigma\beta_j$ and $\Sigma\gamma_j$ be constrained to equal Σw_j, apparently meaning that the "long-run MPC" out of any type of income is identical. However, the PIH-cum-rational-expectations approach suggests no such adding-up constraint. To see this, follow Sargent (1978, pp. 681–82) in rewriting (2) in the form $X_t = HX_{t-1} + \eta_t$ (Sargent's eq. 8), where

$$X_t = \begin{bmatrix} Y_t \\ Y_{t-1} \\ \cdot \\ \cdot \\ \cdot \\ \cdot \\ Y_{t-n} \end{bmatrix}, \quad H = \begin{bmatrix} a_1 & a_2 & \ldots & a_n & a_{n+1} \\ 1 & 0 & & 0 & 0 \\ 0 & 1 & & \cdot & \cdot \\ \cdot & 0 & & \cdot & \cdot \\ \cdot & \cdot & & \cdot & \cdot \\ \cdot & \cdot & & 0 & \cdot \\ 0 & 0 & & 1 & 0 \end{bmatrix}, \quad \eta_t = \begin{bmatrix} \epsilon_t \\ 0 \\ \cdot \\ \cdot \\ \cdot \\ \cdot \\ 0 \end{bmatrix},$$

and $Y_t = dX_t$, where $d = (1, 0, \ldots, 0)$. As Sargent notes (his eq. 9), rational expectations implies ${}_tX_{t+s} = H^sX_t$, whence ${}_tY_{t+s} = dH^sX_t$. Substituting this into (1) gives the following expression for permanent income:

$$Y_t^p = A_t + \left[\sum_{s=0}^{\infty} \frac{dH^s}{(1+r)^s}\right]X_t,$$

which is of the form (4) with coefficients

$$b \equiv (b_0, b_1, \ldots, b_n) = \sum_{s=0}^{\infty} \frac{dH^s}{(1+r)^s}.$$

The sum of the b_j has no obvious interpretation. Thus, in a model with several sources of income, there is no particular reason why the various sets of distributed lag coefficients should have a common sum.

Where, then, does the lifetime budget constraint enter? The answer is that (4) implies a unitary lifetime MPC for any values of k and the b_j. The proof involves some straightforward but tedious algebraic manipulations of the difference equations (4) and

$$A_{t+1} = (1+r)A_t + Y_t - C_t \tag{10}$$

and hence is relegated to Appendix A.

With these preliminaries out of the way, it is easy to explain the estimating equation. If there were no special taxes to worry about, the basic empirical model of consumer behavior would be as follows:

$$\begin{aligned} C_t = {} & k_0 + k_1 r_t Y_t + \sum_{j=0}^{n} w_j Y_{t-j} \\ & + k_2(W_t - A_t) + \sum_{j=0}^{q} k_{3+j} A_{t-j} + u_t, \end{aligned} \tag{11}$$

where r_t is the rate of interest, W_t is consumer net worth at the beginning of period t, and A_t is the market value of stock market wealth at the beginning of period t. The specific way in which assets are entered into the consumption function, including the constraint that $k_3 + k_4 + \ldots + k_{3+q} = k_2$, is suggested by the MIT-Penn-SSRC (MPS) model and is unimportant to what follows.

Now consider the separation of disposable income into its two components:

$$Y_t = R_t + S_t. \tag{12}$$

The way I have defined the γ's means that (11) is expanded to:

$$C_t = k_0 + k_1 r_t Y_t + k_2(W_t - A_t) + \sum_{j=0}^{q} k_{3+j} A_{t-j}$$

$$+ \sum_{j=0}^{n} w_j R_{t-j} + \sum_{j=0}^{n} \gamma_j^1(t) S_{t-j}^1 \tag{13}$$

$$+ \ldots + \sum_{j=0}^{n} \gamma_j^m(t) S_{t-j}^m + u_t.$$

Substituting (9) and (12) into (13), and rearranging terms, gives:

$$C_t = k_0 + k_1 r_t Y_t + k_2(W_t - A_t) + \sum_{j=0}^{q} k_{3+j} A_{t-j}$$

$$+ \sum_{j=0}^{n} w_j (Y_{t-j} + \lambda X_t^j - S_{t-j}) \tag{14}$$

$$+ \sum_{j=0}^{n} \beta_j \left[(1 - \lambda) X_t^j + \sum_{i=1}^{m} (1 - D_t^i) S_{t-j}^i \right] + u_t,$$

where $X_t^j = D_t^1 S_{t-j}^1 + \ldots + D_t^m S_{t-j}^m$, $j = 0, \ldots, n$. This is not the actual estimating equation because additional income sources were distinguished, because the distributed lag coefficients were constrained in several ways, and because corrections were made for both heteroscedasticity and serial correlation in the error term. Details are spelled out in Appendix B. Nonetheless (14) is the most useful form for interpreting the estimated parameters. The model is nonlinear because of the parameter λ—the crucial parameter of this study.

As noted above, theory does not imply that $\Sigma w_j = \Sigma \beta_j$. To investigate this further, this adding-up constraint was imposed in (14) and its validity tested as follows.[10] Let ℓ denote the likelihood ratio. Then, under the assumption of normality, $-2 \log \ell = T \log (\mathrm{SSR}_r/\mathrm{SSR})$ is distributed as a χ^2 with r degrees of freedom, where: T = number of observations (= 100 throughout this paper), SSR = minimized value of the sum of squared residuals in the unconstrained regression (eq. [14] in this case), SSR_r = minimized value of the sum of squared residuals in the constrained regression (obtained by imposing $\Sigma\beta = \Sigma w$), and r = number of restrictions (= 1 in this case). As reported in table 1, row 1, the constraint was rejected at the 10 percent level but not at the 5 percent level. There being no persuasive theoretical rationale for it, the constraint was dropped.

Tacitly, however, (14) embodies a number of other constraints that are equally lacking in theoretical justification—constraints that each type of regular income is subject to the same set of distributed lag coefficients. These constraints were tested by a series of likelihood ratio tests, which are described in the balance of this section.

[10] See Goldfeld and Quandt (1972, p. 74).

TABLE I

χ^2 TESTS OF CONSTRAINTS

Unconstrained Model	Constraint Tested	df	Test Statistic
1. Eq. (14)	$\Sigma w = \Sigma \beta$	1	3.67
2. Eq. (16)	$v_j = w_j$ for all j	3	12.02
3. Eq. (16)	$\Sigma v_j = \Sigma w_j$	1	Approximately 0*
4. Eq. (18) with $\Sigma v = \Sigma w$	$\phi_j = w_j$ for all j	3	19.02
5. Eq. (18) with $\Sigma v = \Sigma w$	$\Sigma \phi_j = \Sigma w_j$	1	12.52
6. Eq. (18) with $\Sigma v = \Sigma w$	$\lambda = 0$	1	1.32
7. Eq. (18) with $\Sigma v = \Sigma w$	$\lambda = 1$	1	2.84
8. Eq. (18) with $\Sigma v = \Sigma w$	$k_2 = 0$	1	17.81

NOTE.—Critical levels for the χ^2 distribution are:

df	10% Point	5% Point	1% Point
1	2.71	3.84	6.63
3	6.25	7.81	11.34

*Due to rounding error the actual computed test statistic was slightly negative. When this constraint was tested in the context of eq. (18), it produced a test statistic of 1.32.

First, "regular" disposable income was disaggregated into its two main components—personal income and "regular" personal taxes:

$$R_t = P_t - T_t.^{11} \tag{15}$$

Personal income is assumed to be spent according to the lag weights w_j, while regular taxes are assumed to be subject to a different set of lag weights v_j. Special taxes are treated as previously explained, except that the v's replace the w's in equations (8) and (9). That is, while the tax is on the books, a special tax is treated as a weighted average of a permanent tax and a windfall. Thus the basic consumption function becomes:

$$\begin{aligned} C_t = {} & k_0 + k_1 r_t Y_t + k_2(W_t - A_t) + \sum_{j=0}^{q} k_{3+j} A_{t-j} \\ & + \sum_{j=0}^{n} w_j P_{t-j} - \sum_{j=0}^{n} v_j (T_{t-j} - \lambda X_t^j) \\ & + \sum_{j=0}^{n} \beta_j \left[(1-\lambda) X_t^j + \sum_{i=1}^{m} (1 - D_t^i) S_{t-j}^i \right] + u_t. \end{aligned} \tag{16}$$

Since (as explained below) the distributed lag coefficients are constrained to follow a third-degree polynomial with a zero end-point

[11] In making this separation, I departed a bit from national income accounting conventions by including the employer's share of social insurance contributions in both P and T.

constraint, the null hypothesis that the v_j are equal to the w_j imposes three constraints on equation (16). Row 2 of table 1 shows that these constraints were resoundingly rejected by the data ($\chi_3^2 = 12$). However, it turned out that the sum of the v_j was estimated to be almost exactly equal to the sum of the w_j (see row 3 of table 1), so this adding-up constraint was imposed in subsequent estimates.

The final generalization considered was to disaggregate personal income into its two main components—factor income and transfer payments:[12]

$$P_t = F_t + V_t. \tag{17}$$

This made the estimating equation:

$$\begin{aligned} C_t = k_0 + k_1 r_t Y_t + k_2(W_t - A_t) + \sum_{j=0}^{q} k_{3+j} A_{t-j} \\ + \sum_{j=0}^{n} w_j F_{t-j} + \sum_{j=0}^{n} \phi_j V_{t-j} - \sum_{j=0}^{n} v_j (T_{t-j} - \lambda X_t^j) \\ + \sum_{j=0}^{n} \beta_j \left[(1 - \lambda) X_t^j + \sum_{i=1}^{m} (1 - D_t^i) S_{t-j}^i \right] + u_t. \end{aligned} \tag{18}$$

Once again, the null hypothesis that the ϕ_j (spending coefficients for transfers) are in fact equal to the w_j (spending coefficients for factor income) was tested by a χ^2 test. And once again it was resoundingly rejected ($\chi_3^2 = 19$; see row 4 of table 1). This time, however, the data also rejected the adding-up constraint $\Sigma\phi_j = \Sigma w_j$, which therefore was not imposed (table 1, row 5). In fact, most of the difference between the ϕ_j and the w_j was in their sums; the time patterns were remarkably similar.

To summarize these tests, we are left with a model that assigns distributed lag weights w_j to factor income, ϕ_j to transfers, v_j to regular taxes, and β_j to windfalls. The Σv_j and Σw_j are apparently equal, but the other sums are not.

V. Issues in Estimation

1. *Data*

Following the suggestion of Darby (1975), I used consumer expenditures, rather than pure consumption, as C_t. This seems most appropriate where the focus is on the evaluation of stabilization policy, as it is here, rather than on testing the PIH. Furthermore, it

[12] For this purpose, both employer contributions and business transfers are considered to be factor income, and the aspects of the 1975–76 tax cuts that are classified as transfer payments in the national income accounts are grouped with temporary taxes.

TABLE 2

EFFECTS ON DISPOSABLE INCOME OF 1975–76 TAX CUTS AND TRANSFERS*

	TAXES†		TRANSFERS‡		
QUARTER	Tax Rate Cuts	Rebate	Social Security	Earned Income	TOTAL
1975:2	8.5	31.2	6.7	0	46.4
1975:3	12.3	0	0	0	12.3
1975:4	11.9	0	0	0	11.9
1976:1	14.2	0	0	1.9	16.1
1976:2	14.0	0	0	1.6	15.6

* In billions of current dollars, at annual rates
† From U.S. Bureau of Economic Analysis (February 1976, March 1977)
‡ Kindly supplied to the author by Joseph C. Wakefield, of the Bureau of Economic Analysis, in conversation

avoids many complicated issues of definition (e.g., Which goods are durables? How fast do they depreciate? etc., etc.). The cost of this shortcut is that the theoretical interpretation of some of the parameters is lost. For example, k_1 includes the effects of r_t on both pure consumption (which may be positive or negative) and spending on durables (which should be negative). Similarly, the lag weights w_j might be expected to be less smooth than the lag of pure consumption behind income because of the lumpy nature of durables.[13]

Data for the 1968 surcharge were taken from Okun (1971) and converted to 1972 dollars by the deflator for personal consumer expenditures; these comprise S_t^1. Data for the various 1975–76 tax cuts are shown in table 2; they were similarly deflated and segregated into two time series. The S_t^2 series is defined as the explicitly one-shot measures: the tax rebates and the social security bonuses (hereafter referred to as "the rebate"). The rest is considered as S_t^3. Since the 1975 cuts were extended several times and are now a permanent feature of the tax code, an arbitrary decision had to be made as to when they became "permanent."[14] I decided to cut off S_t^3 after 1976:2, because by then the cuts had already been extended once in the Revenue Adjustment Act of 1975 and a second time in the Tax Reform Act of 1976.

[13] To account for the special features of expenditures on consumer durables, several additional variables were tried in some earlier regressions. Neither the stock of durables, nor the relative price of durables, nor the unemployment rate succeeded in significantly lowering the sum of squared residuals. In several cases, the signs of the coefficients were even the opposite of what theory suggests. It may be that these variables are more relevant to the choice between saving in the form of durables vs. in financial form than they are to the choice between spending and saving.

[14] At an early stage of this research, I experimented with a learning model in which a temporary tax gradually came to be considered permanent as it remained on the books longer and longer. This experiment was unsuccessful.

Data on consumer net worth (W_t), and its breakdown into stock market (A_t) and non–stock market ($W_t - A_t$) components, were taken from the data bank of the MPS model and converted to 1972 dollars. They are based on a number of primary sources, the most important of which is the flow of funds.

Because of the recent findings of Boskin (1978), I thought it important to use a real after-tax interest rate. Since the construction of this series was a fairly involved affair, I explain it only in Appendix C.

2. *Distributed Lag Estimation*

The many distributed lags in equation (18) were estimated by an adaptation of the Almon (1965) lag technique, as a method of conserving on parameters. Generally, a third-degree polynomial with a zero constraint at the far end was used. Preliminary tests suggested that these end-point constraints ($w_{n+1} = 0$, etc.) could not be rejected. An unconstrained version of one specification, run as an experiment, showed that the polynomial constraint had very little effect on the w_j coefficients but did influence the β_j coefficients.

The distributed lag effects of assets were estimated as follows. In some preliminary regressions, the two components were combined and a cubic distributed lag over n quarters was estimated. Then the two components were disaggregated. These preliminary tests showed quite clearly (*a*) that the coefficients were not the same and (*b*) that the lag was much shorter than n quarters. (As explained just below, n was chosen to be 7.) In the case of non–stock market wealth, an estimated distributed lag over 4 quarters attached virtually all of the weight to the current (start of period) value, so all lagged values were omitted. In the case of stock market wealth, when a cubic was estimated over $j = 0, 1, \ldots, 7$, the coefficients turned out to be almost linear and to be virtually zero after $j = 2$ (a small positive value for $j = 3$ and small negative values for $j = 4, \ldots, 7$). Thus a linear distributed lag over $j = 0, \ldots, 3$ was adopted as the final specification.

The length of the distributed lag, n, was selected by running a preliminary version of the regression for alternative values of n ranging from 6 to 10. A very clear minimum in the sum of squared residuals was found around $n = 7$ or $n = 8$, with the former having a slightly better fit and slightly better coefficients. On this basis, $n = 7$ was selected for all subsequent work.

3. *Treatment of the 1975 Rebate*

As has been noted already, the model divides the 1975–76 tax cuts into two parts: S_t^2 includes the rebate, while S_t^3 includes all the rest.

This makes a strong (and questionable) assumption about how consumers treated the rebate. In particular, it assumes that they treated it just like the 1968 surcharge and the other 1975 reductions: essentially *as if a fraction λ of it was a regular increase in income, while a fraction $1 - \lambda$ was a pure windfall. Given the nature of the rebate, this is questionable, to say the least.*

An alternative assumption—equally strong as the first—is that consumers treated the rebate as a pure windfall right from the start.[15] While it is not obvious that this is true, since consumers might have anticipated a repeat performance with some reasonable probability, it does seem a plausible working hypothesis. Fortunately, it is not difficult to modify any of the three models to accommodate this alternative hypothesis; all that is necessary is that the lag weights β_j be applied to S_t^2 starting immediately in 1975:2. When this was done in one version of the model, the resulting equation had virtually an identical SSR, almost the same estimated λ, and very similar implications about spending patterns out of the rebate. Thus the conclusions of this study seem insensitive to the treatment of the rebate.

Accession Number: 67790
Date 30.1.83

VI. Empirical Results

1. *Parameter Estimates and Interpretation*

Estimation was done by the numerical optimization package developed by S. M. Goldfeld and R. E. Quandt. The results from estimating equation (18) on quarterly data covering 1953:1–1977:4 are presented in table 3. The number in parentheses next to each estimated coefficient is its asymptotic standard error (or rather a numerical estimate thereof).

In interpreting the standard error of the regression, it should be mentioned that, as explained in Appendix B, the equation was actually transformed so that the left-hand variable was the average propensity to consume (APC), C_t/Y_t, rather than consumer spending. Thus, the standard error of .0038 is relative to a typical value for the APC of about .90. This represents an excellent fit.[16] At 1977 income levels, it translates to a standard error of about $3.5 billion in predicting consumer spending. Of course, obtaining a good fit with a consumption function is hardly a notable achievement, and the equation—like most consumption functions—does suffer from some autocorrelation (ρ = .54).

[15] Since the rebate was off the books by 1975:3, and hence treated as a windfall in any case, only 1975:2 is at issue here.

[16] The standard errors of a comparable equation in Modigliani and Steindel (1977) are .0056 with an autocorrelation correction and .0065 without.

TABLE 3

NONLINEAR CONSUMPTION FUNCTION PARAMETER ESTIMATES*

k_0	k_1	k_2	λ	ρ
30.1	.0002	.021	.50	.54
(7.1)	(.0007)	(.005)	(.32)	(.11)

DISTRIBUTED LAG COEFFICIENTS†

j	w_j	ϕ_j	v_j	β_j	k_{3+j}	γ_j	Rebate
0	.60 (.05)	.50 (.16)	.34 (.09)	−.03 (.26)	.009 (.002)	.16	.16
1	.16 (.02)	−.02 (.09)	.14 (.04)	−.03 (.11)	.006 (.001)	.06	−.03
2	−.06 (.03)	−.20 (.10)	.04 (.05)	−.01 (.09)	.004 (.001)	.02	−.01
3	−.12 (.02)	−.14 (.08)	.01 (.05)	.03 (.09)	.002 (.0005)	.02	.03
4	−.08 (.01)	.04 (.06)	.02 (.03)	.08 (.07)	..	.05	.08
5	.02 (.01)	.26 (.07)	.05 (.03)	.11 (.06)	...	.08	.11
6	.10 (.02)	.39 (.09)	.08 (.04)	.12 (.08)	..	.10	.12
7	.11 (.02)	.34 (.08)	.07 (.04)	.09 (.07)	..	.08	.09
Sum	.74	1.17	.74	.36	.021 (.005)	.55	.55

NOTE.—k_0 = constant, k_1 = coefficient of interest rate, k_2 = coefficient of non-stock market wealth, λ = weight on regular income, ρ = autocorrelation coefficient. Sum of squared residuals = .00147, SE = .00383, SE of unadjusted errors (*without a correction for autocorrelation*) = .00455, N observations = 100.

* Asymptotic SEs are in parentheses.

† Components may not add to totals due to rounding.

Turning to the coefficient estimates, the most critical parameter for purposes of this study is λ, the weight attached to regular income in equation (8). The point estimate of .50 suggests that temporary taxes that are still on the books are treated like 50-50 blends of windfalls and regular taxes. However, the standard error is regrettably large; there are, after all, pitifully few observations that can be used to estimate λ. The null hypotheses $\lambda = 0$ or $\lambda = 1$ can nonetheless be tested by likelihood ratio tests. When these tests were run with equation (18) as the unconstrained regression (see table 1, rows 6 and 7), the null hypothesis that $\lambda = 0$ (temporary taxes are regarded as pure windfalls) could not be rejected ($\chi_1^2 = 1.32$). But the null hypothesis that $\lambda = 1$ (temporary taxes are regarded as regular income) could be rejected if we were not too fussy about significance levels ($\chi_1^2 = 2.84$).

I turn next to the distributed lag coefficients of the various income terms. The w's for factor income are very large and positive at first, then turn small and negative, and finally become positive again at the end. This general shape accords well with our expectations.[17] The ϕ_j coefficients for transfer payments follow a similar shape but are much more erratic and less well pinned down econometrically. A notable

[17] Because of the estimating form, it is unlikely that simultaneity has much to do with the large estimate for w_0. See Appendix B.

feature is that their sum is nearly 1.2, indicating "overspending" during the first 2 years after receipt of a transfer payment.

The v_j coefficients for regular taxes also exhibit a characteristic U-shape, but in much more muted fashion. As compared with factor income, spending in the first year after a regular tax cut is apparently substantially less, after which it catches up. This was surprising at first, since Dolde (1976) and Modigliani and Steindel (1977) had suggested that regular tax changes are "more permanent" than regular income.[18] However, it turns out that the following second-order autoregressives describe the deviations from trend of personal income (P_t) and regular taxes (T_t):

$$P_t = \underset{(.09)}{1.32P_{t-1}} - \underset{(.09)}{.38P_{t-2}},$$

$$T_t = \underset{(.09)}{1.11T_{t-1}} - \underset{(.10)}{.21T_{t-2}}.$$

Using the formulas derived earlier for permanent income, these time-series models imply that "permanent personal income" and "permanent taxes" are given by $P_t^p = 14.9P_t - 5.7P_{t-1}$ and $T_t^p = 9.3T_t - 1.9T_{t-1}$, so that, contrary to Dolde and Modigliani and Steindel, we should actually expect a stronger short-run response of consumption to fluctuations in P_t than to fluctuations in T_t, which is exactly what I find.

The next column, the β_j's, are in some sense out-of-sample extrapolations since there are no "pure windfalls" recorded in the data.[19] Their only use is to form the weighted average $\lambda w_j + (1 - \lambda)\beta_j$, which is reported in the column marked "γ_j." These are the expenditure coefficients for income from a temporary tax that remains on the books for the entire 2-year horizon. To illustrate the opposite extreme, the column marked "Rebate" shows the spending coefficients for a temporary tax that lasts only 1 quarter. These two columns differ in details but are quite similar. There is a moderate spending response in the initial quarter, followed by very little spending over the next 3 or 4 quarters. Most of the spending out of a temporary tax cut, according to these estimates, comes 5 or more quarters after the cut.

The coefficient of assets (.02) is comparable to what others have estimated, though a bit on the low side. A likelihood ratio test of the null hypothesis $k_2 = 0$ (which, in this constrained form, also implies k_3

[18] But see Dolde (1979), where the transitory nature of allegedly permanent tax changes is stressed

[19] I made an attempt, in some early regressions, to treat the National Service Life Insurance Dividends of 1950 in this way, but I was not successful.

TABLE 6

ESTIMATED EFFECTS OF THE 1975–76 TEMPORARY TAX CUTS ON CONSUMER EXPENDITURES*

Quarter	Estimated Spending Effect
1975:2	5.9
1975:3	1.6
1975:4	2.7
1976:1	4.8
1976:2	6.7
1976:3	6.7
1976:4	7.1
1977:1	6.6
1977:2	7.2
1977:3	7.5
1977:4	7.3

* In billions of 1972 dollars

percent—the lowest figure that had then been recorded since the Korean War.)

According to the estimates presented in this paper, the temporary tax cuts of 1975–76 contributed to both phenomena. Using the spending coefficients presented in table 4, table 6 shows the estimated direct (excluding multiplier) effects on consumer spending of the tax cuts of 1975:2 through 1976:2, inclusive.[21] It appears that (1) very little of the rebate was spent in 1975:2, (2) rather little of the disposable income attributable to the temporary tax cut package was spent during the remainder of 1975, and (3) more spending out of the temporary tax cuts was done in 1976 and yet more in 1977. Both the low APC of 1975:2 and the high APCs of late 1976 and early 1977 are tracked very well by the model. One observation which virtually jumps from table 6 is how very small these estimated spending impacts are relative to the size of the economy they were meant to stimulate (real GNP in the neighborhood of $1,250 billion).

Finally, there is one more question. If, instead of the 1975:2–1976:2 package of temporary measures, the government had cut taxes "permanently" in 1975:2 and then restored them to their original level starting in 1976:3, how large a tax cut would have achieved the same average effect on aggregate demand?[22]

Table 7 summarizes the model's answers to this question for three different choices of the horizon over which the "average effect on aggregate demand" might be defined. The first column gives the average direct impact on spending attributed by the model to the

[21] The reader is reminded that only these 5 quarters are considered temporary cuts.

[22] For this calculation I assume that consumers were successfully fooled into thinking that the 5-quarter tax cut would be permanent.

TABLE 7

EQUIVALENT PERMANENT TAXES*,†

Horizon	Average Impact on Spending of Actual 1975–76 Cuts	Cumulative Revenue Loss over 5 Quarters of Permanent Tax Cut with Equal Average Impact on Spending
4 quarters	3.7	9.5
6 quarters	4.7	12.4
8 quarters	5.3	15.9

* See text for definition
† In billions of 1972 dollars, at annual rates

1975–76 tax cuts. The next column shows how much total tax revenue the government would have had to relinquish during the same 5-quarter period (1975:2–1976:2) in order to achieve the same direct impact on spending through a permanent tax cut. Since the total 5-quarter revenue loss from the 1975–76 package was $20 billion, these numbers mean, for example, that a permanent tax cut about half as large ($9.5 billion vs. $20 billion) would have had the same first-year effect on aggregate demand. Over a 2-year horizon, however, the 1975–76 package had about 80 percent as much "bang for the buck" as a permanent tax cut.

VII. Summary

Both economic theory and casual empirical observation of the U.S. economy suggest that short-run spending propensities from temporary tax changes are smaller than those from permanent ones, but neither provides much guidance about the magnitude of this difference. This paper offers new empirical estimates of this difference and finds it to be quite substantial.

The analysis is based on an amendment of the standard distributed lag version of the PIH that distinguishes temporary taxes from other income on the grounds that the latter is "more transitory." This amendment, which is broadly consistent with rational expectations, leads to a nonlinear consumption function.

Though the standard error is unavoidably large, the point estimate suggests that a temporary tax change is treated as a 50-50 blend of a normal income tax change and a pure windfall. Over a 1-year planning horizon, a temporary tax change is estimated to have only a little more than half as much impact as a permanent tax change of equal magnitude, and a rebate is estimated to have only about 38 percent as much impact. The model tracks both the extraordinarily high savings rate of 1975:2 and the extraordinarily low savings rates of late 1976 and early 1977 very well and attributes part of both phenomena to the temporary tax measures of 1975–76. Finally, it is estimated that a

permanent tax cut of about \$9.5 billion (in 1972 dollars) would have had the same impact on aggregate demand over the first 4 quarters as the \$20 billion of 1975–76 tax cuts.

Appendix A

The Lifetime Budget Constraint

This Appendix demonstrates a result that, to my knowledge, is not very well known: that the long-run MPC corresponding to a consumption function of the form

$$C_t = kA_t + \sum_{j=0}^{n} b_j Y_{t-j} \tag{A1}$$

is unity for any value of k (greater than the real rate of interest) and for any b's.

Proof: Write equation (10) in the text as

$$[1 - (1 + r)L]A_t = Y_{t-1} - C_{t-1}, \tag{A2}$$

and write (A1) as

$$C_t = kA_t + b(L)Y_t, \tag{A3}$$

where L is the lag operator and $b(L) = b_0 + b_1L + \ldots + b_nL^n$. Applying the operator $1 - (1 + r)L$ to (A3) and using (A2) yields $[1 - (1 + r - k)L]C_t = \{kL + [1 - (1 + r)L]b(L)\}Y_t$, which can be written

$$C_t = B(L)Y_t, \tag{A4}$$

where

$$B(L) = \frac{kL + [1 - (1 + r)L]b(L)}{1 - (1 + r - k)L}. \tag{A5}$$

To obtain the lifetime spending generated by a \$1.00 impulse in Y_t, we must compute the discounted sum of coefficients:

$$\text{MPC} = \sum_{i=0}^{\infty} \frac{B_i}{(1 + r)^i}. \tag{A6}$$

To simplify the notation, let $\theta \equiv 1 + r - k$. Assuming that $\theta < 1$ (i.e., that $k > r$), (A5) can be written: $B(L) = \{kL + [1 - (1 + r)L]b(L)\}(1 + \theta L + \theta^2L^2 + \ldots)$. This is of the form $B(L) = B_0 + B_1L + B_2L^2 + \ldots$, with the following coefficients:

$$B_0 = b_0,$$

$$B_1 = \theta b_0 + [b_1 + k - (1 + r)b_0],$$

$$B_2 = \theta^2 b_0 + \theta[b_1 + k - (1 + r)b_0] + [b_2 - (1 + r)b_1],$$

$$\vdots$$

$$B_n = \theta^n b_0 + \theta^{n-1}[b_1 + k - (1 + r)b_0] + \ldots + [b_n - (1 + r)b_{n-1}],$$

$$B_{n+1} = \theta^{n+1}b_0 + \theta^n[b_1 + k - (1 + r)b_0] + \ldots + \theta[b_n - (1 + r)b_{n-1}] - (1 + r)b_n,$$

$$B_{n+1+s} = \theta^s B_{n+1}, \qquad s = 1, 2, \ldots.$$

Substitution of all of these into (A6), and some truly horrendous grinding, establishes that MPC = 1 regardless of the magnitudes of k and the b_j (as long as $k > 0$). Q.E.D.

A brief word on the interpretation of the sum of the b_j in (A1) may be in order here. Should Y_t rise permanently by \$1.00—a statement that is basically meaningless if the autoregressive process assumed in the text (eq. [2]) really holds—the eventual change in C_t would, by (A4), be $B(1) = \Sigma_{t=0}^{\infty} B_t$. According to (A5), this sum is

$$B(1) = \frac{k - r\sum_{j=0} b_j}{k - r} = 1 + \frac{r}{k - r}\left(1 - \sum_{j=0}^{n} b_j\right).$$

Thus Σb_j controls the size of the spending response to a hypothetical permanent rise in income; it does not influence the lifetime MPC.

Appendix B

Details on the Estimating Equation

This Appendix derives and explains the equation that was actually estimated. I begin by repeating equation (14) of the text:

$$C_t = k_0 + k_1 r_t Y_t + k_2(W_t - A_t) + \sum_{j=0}^{q} k_{3+j} A_{t-j}$$
$$+ \sum_{j=0}^{n} w_j(Y_{t-j} + \lambda X_t^j - S_{t-j}) + \sum_{j=0}^{n} \beta_j\left[(1 - \lambda)X_t^j + \sum_{i=1}^{m} (1 - D_t^i)S_{t-j}^i\right] + u_t.$$

For purposes of reducing heteroscedasticity, the assumption was made that the standard deviation of u_t was proportional to Y_t, so the whole equation was divided through by Y_t to get

$$\begin{aligned} \mathrm{APC}_t = \frac{k_0}{Y_t} &+ k_1 r_t + k_2\left(\frac{W_t}{Y_t} - \frac{A_t}{Y_t}\right) + \sum_{j=0}^{q} k_{3+j}\frac{A_{t-j}}{Y_t} \\ &+ \sum_{j=0}^{n} w_j(z_{t-j} - q_{t-j}) + \sum_{j=0}^{n} \beta_j\left[(1 - \lambda)x_t^j + \sum_{i=1}^{m}(1 - D_t^i)s_{t-j}^i\right] + \epsilon_t, \end{aligned} \tag{B1}$$

where

$z_{t-j} \equiv Y_{t-j}/Y_t$ (note: $z_t = 1$ for all t),

$x_t^j = X_t^j/Y_t$,

$s_{t-j} = \dfrac{S_{t-j}}{Y_t}$,

$q_{t-j} = s_{t-j} - \lambda x_t^j$,

$\epsilon_t = \dfrac{u_t}{Y_t}$.

To estimate (B1), the assumption was made that both w_j and β_j follow third-degree polynomials in j:

$$w_j = a_0 + a_1 j + a_2 j^2 + a_3 j^3,$$
$$\beta_j = b_0 + b_1 j + b_2 j^2 + b_3 j^3. \tag{B2}$$

The end-point constraints mentioned in the text ($w_{n+1} = \beta_{n+1} = 0$) are thus

$$a_0 + (n+1)a_1 + (n+1)^2 a_2 + (n+1)^3 a_3 = 0,$$
$$b_0 + (n+1)b_1 + (n+1)^2 b_2 + (n+1)^3 b_3 = 0. \tag{B3}$$

Equations (B3) were used to eliminate the parameters a_0 and b_0. The adding-up constraint discussed (and rejected) in the text was thus

$$\sum_{j=0}^{n} \{a_1[j - (n+1)] + a_2[j^2 - (n+1)^2] + a_3[j^3 - (n+1)^3]\},$$
$$= \sum_{j=0}^{n} \{b_1[j - (n+1)] + b_2[j^2 - (n+1)^2] + b_3[j^3 - (n+1)^3]\}, \tag{B4}$$

which was used to eliminate the parameter a_1.

Finally, in estimating (B1), ϵ_t was assumed to follow a first-order autoregressive scheme, $\epsilon_t = \rho\epsilon_{t-1} + e_t$, where e_t is white noise. Estimation was by nonlinear least squares, which is equivalent to maximum likelihood if e_t is normally distributed. The function actually minimized was

$$\sum_{t=1}^{T} \Bigg\{ \mathrm{APC}_t - \frac{k_0}{Y_t} - k_1 r_t - k_2\left(\frac{W_t - A_t}{Y_t}\right)$$
$$- \sum_{j=0}^{n} k_{3+j} \frac{A_{t-j}}{Y_t} - \sum_{j=0}^{n} w_j(z_{t-j} - q_{t-j})$$
$$- \sum_{j=0}^{n} \beta_j \left[(1-\lambda)x_t^j - \sum_{i=1}^{m} (1 - D_t^i)s_{t-j}^i\right] - \rho\epsilon_{t-1} \Bigg\}^2,$$

with all the above-mentioned definitions and parameter restrictions substituted in.

The estimating forms when regular income was further disaggregated were derived in precisely analogous ways from equations (16) and (18).

Appendix C

Calculation of the Real After-Tax Interest Rate

The real after-tax interest rate is defined as $r_t = i_t(1 - \tau_t) - \pi_t$, where i_t is the nominal interest rate, τ_t is the marginal tax rate, and π_t is the expected rate of inflation.

Nominal interest rate.—Four different nominal interest rates were tried: a corporate bond rate, the 4–6 month commercial paper rate, the 3-month Treasury bill rate, and a weighted average of rates received on various time and savings accounts. While all four led to very similar results, the Treasury bill rate gave the best fit and so was adopted.

Marginal tax rate.—The marginal tax rate was created from the average tax rate in the following way. Let $T(y)$ be the tax function facing an individual

taxpayer, and let $f(y)$ be the frequency distribution of income. The only directly observable tax rate is the aggregate average rate, which is $A = [\int T(y)f(y)dy]/[\int yf(y)dy]$.

Now suppose the whole income distribution shifts to the right with no change in its shape; that is, it shifts from $f(y)$ to $f[y(1-h)]$, where h connotes a "small" multiplicative shift. The average income is then $Y(h) = \int yf[y(1-h)]dy$, so that, for small shifts (h near zero), $dY/dh = -\int y^2f'(y)dy$. Similarly, average tax payments after the shift are $T(h) = \int T(y)f[y(1-h)]dy$, so that $dT/dh = -\int yT(y)f'(y)dy$. The aggregate marginal tax rate, M, is the ratio of these: $M \equiv [\int yT(y)f'(y)dy]/[\int y^2f'(y)dy]$.

To evaluate these integrals and obtain a closed expression for M/A, I adopted the following functional forms:

$$T(y) = ay^b \qquad (b > 1,\ a > 0),$$
$$f(y) = \gamma e^{-\gamma y} \qquad (\gamma > 0).$$

With these assumptions, the two ratios of integrals work out to be $A = (a\gamma^2/\gamma^{b+1})\Gamma(b+1)$, $M = (a\gamma^3/2\gamma^{b+2})\Gamma(b+2)$, where $\Gamma(n)$ is the "gamma function," namely, $\Gamma(n+1) = n\Gamma(n)$. Hence $M/A = (b+1)/2$, and using a tax elasticity of $b = 1.6$ gives $M = (1.3)A$.

The average tax rate was computed, quarter by quarter, by dividing the sum of federal and state-local personal tax and nontax payments by personal income excluding transfers (an approximation to the tax base). These are all official national income accounts series.

Expected rate of inflation.—Inflationary expectations were generated by a model based on what has been called "economically rational" expectations.[23] The idea is that agents, in informing themselves, begin with the data that are cheapest per unit of informational content and then proceed to process more costly data until the marginal cost and (expected) marginal benefits are equated. In this particular application, I assumed that consumers base their expectations of the inflation rate ($\dot{P}_t$) on its own past history and on the history of the growth rate of the money supply ($\dot{M}_t$). Thus I estimated an equation

$$\dot{P}_t = a_0 + \sum_{j=1}^{J} a_j\dot{P}_{t-j} + \sum_{i=0}^{I} b_i\dot{M}_{t-i} + e_t$$

on actual U.S. quarterly data, using the deflator for personal consumption expenditures for P_t and M_2 for M_t, and assumed that consumers used this equation to generate expectations. In estimation, I used the Almon (1965) lag technique with third-degree polynomials, no end-point constraints, and various choices for J and I. The best results were obtained with $J = 11$ and $I = 17$, namely[24]

$$P_t = \underset{(.38)}{-.60} + \sum_{1}^{11} a_jP_{t-j} + \sum_{0}^{17} b_iM_{t-i},\ R^2 = .73, \qquad \text{D-W} = 1.98,$$

standard error = 1.42, mean of dependent variable = 3.35,

$$\Sigma a_j = \underset{(.10)}{.67} \qquad \Sigma b_i = \underset{(.09)}{.27}.$$

[23] See Feige and Pearce (1976).

[24] Standard errors are in parentheses. The D-W is the Durbin-Watson statistic. The period of estimation was 1951:3–1977:4, the longest period possible given the need for 17 lagged values of M.

References

Almon, Shirley. "The Distributed Lag between Capital Appropriations and Expenditures." *Econometrica* 33 (January 1965): 178–96.

Blinder, Alan S. "Distribution Effects and the Aggregate Consumption Function." *J.P.E.* 83, no. 3 (June 1975): 447–75.

———. "Intergenerational Transfers and Life Cycle Consumption." *A.E.R. Papers and Proc.* 66 (May 1976): 87–93.

Blinder, Alan S., and Solow, Robert M. "Analytical Foundations of Fiscal Policy." In *The Economics of Public Finance*, by Alan S. Blinder and others. Washington: Brookings Inst., 1974.

Boskin, Michael J. "Taxation, Saving, and the Rate of Interest." *J.P.E.* 86, no. 2, pt. 2 (April 1978): S3–S27.

Darby, Michael R. "The Allocation of Transitory Income among Consumers' Assets." *A.E.R.* 62 (December 1972): 928–41.

———. "Postwar U.S. Consumption, Consumer Expenditures, and Saving." *A.E.R. Papers and Proc.* 65 (May 1975). 217–22.

Dolde, Walter. "Forecasting the Consumption Effects of Stabilization Policies." *Internat. Econ. Rev.* 17 (June 1976): 431–46.

———. "Capital Markets and the Short Run Behavior of Life Cycle Savers." *J. Finance* 33 (May 1978): 413–28.

———. "Temporary Taxes as Macro-economic Stabilizers." *A.E.R. Papers and Proc.* 69 (May 1979): 81–85.

Eisner, Robert. "Fiscal and Monetary Policy Reconsidered." *A.E.R.* 59 (December 1969): 897–905.

Feige, Edgar L., and Pearce, Douglas K. "Economically Rational Expectations: Are Innovations in the Rate of Inflation Independent of Innovations in Measures of Monetary and Fiscal Policy?" *J.P.E.* 84, no. 3 (June 1976): 499–522.

Foley, Duncan K., and Hellwig, Martin F. "Asset Management with Trading Uncertainty." *Rev. Econ. Studies* 42 (July 1975). 327–46.

Friedman, Milton. *A Theory of the Consumption Function.* Princeton, N.J.: Princeton Univ. Press (for Nat. Bur. Econ. Res.), 1957.

Goldfeld, Stephen M., and Quandt, Richard E. *Nonlinear Methods in Econometrics.* Amsterdam: North-Holland, 1972.

Hall, Robert E. "Stochastic Implications of the Life Cycle–Permanent Income Hypothesis. Theory and Evidence." *J.P.E.* 86, no. 6 (December 1978): 971–87.

Houthakker, Hendrik S., and Taylor, Lester D. *Consumer Demand in the United States.* Cambridge, Mass.: Harvard Univ. Press, 1966.

Juster, F. Thomas. "A Note on Prospective 1977 Tax-Cuts and Consumer Spending." Unpublished paper, Univ. Michigan, January 1977.

Lucas, Robert E., Jr. "Econometric Policy Evaluation. A Critique." In *The Phillips Curve and Labor Markets*, edited by Karl Brunner and Allan H. Meltzer. Amsterdam: North-Holland, 1976.

Modigliani, Franco, and Brumberg, Richard E. "Utility Analysis and the Consumption Function: An Interpretation of Cross-Section Data." In *Post-Keynesian Economics*, edited by Kenneth K. Kurihara. New Brunswick, N.J.: Rutgers Univ. Press, 1954.

Modigliani, Franco, and Steindel, Charles. "Is a Tax Rebate an Effective Tool for Stabilization Policy?" *Brookings Papers Econ. Activity*, no. 1 (1977), pp. 175–203.

Muth, John F. "Optimal Properties of Exponentially Weighted Forecasts." *J. American Statis. Assoc.* 55 (June 1960): 299–306.

Okun, Arthur M. "The Personal Tax Surcharge and Consumer Demand, 1968–70." *Brookings Papers Econ. Activity*, no. 1 (1971), pp. 167–211.

Sargent, Thomas J. "Rational Expectations, Econometric Exogeneity, and Consumption." *J.P.E.* 86, no. 4 (August 1978): 673–700.

Springer, William L. "Did the 1968 Surcharge Really Work?" *A.E.R.* 65 (September 1975): 644–59.

U.S. Bureau of Economic Analysis. *Survey of Current Business*. Washington: Government Printing Office, various issues.

Consumer Search with Uncertain Product Quality

John D. Hey

University of York

Chris J. McKenna

University College, Cardiff

This paper presents a unified model of consumer search when both prices and qualities are uncertain. It is assumed that price can be observed prior to purchase but that quality can be observed only after purchase and experience. This two-stage decision problem is of particular interest since, in practice, there is usually some connection between quality and price. Of the many interesting implications of this dependence explored in this paper is the result that the usual "reservation-price rule" may no longer be optimal; indeed, the optimal stopping set may take one of several alternative (and more realistic) forms

I. Introduction

The economics of search has made considerable progress in recent years in describing and explaining the actions of economic agents operating in uncertain environments. Our concern in this paper is with the application of this large body of knowledge to the consumer-search problem. In its simplest form this problem examines the behavior of a consumer searching (at a constant cost per observation) over a known distribution of prices for the lowest (economically sensible) price for the purchase of one unit of some good. The

We would like to thank an anonymous referee and George J. Stigler for some very helpful comments on two earlier drafts of this paper. We are also grateful to Louis L. Wilde for useful comments on the original draft.

Journal of Political Economy, 1981, vol. 89, no. 1)
© 1981 by The University of Chicago 0022-3808/81/8901-0012$01.50

solution, which is now well known, takes the form of a stopping rule uniquely determined by a "reservation price"; that is, the consumer behaves optimally by searching until a price lower than the reservation price is obtained.

In this simple form, and indeed in many of its more sophisticated generalizations, a crucial assumption is that search takes place over just one characteristic of the good—namely, its price. However, in practice, a good has several dimensions (of which price is but one); moreover, uncertainty surrounds not just the price dimension but these other dimensions as well. Of course, this observation would be trivial if information on these other aspects could be collected at the same time as information about price. However, in practice, some characteristics inevitably remain undiscovered until after purchase; thus, prior to purchase, residual uncertainty remains.

The purpose of this paper is to explore the consumer-search problem in this multidimensional form. In particular, we partition the set of characteristics into those which can be observed prior to purchase and those which can be observed only after purchase. To keep our exposition simple, we consider the simplest of all possible formulations: a two-dimensional case in which *price* is the dimension which can be observed prior to purchase and *quality* is the dimension observable only on purchase and experience.

Our extension makes the consumer's problem a two-stage one: The first stage relates to the price dimension—the consumer must decide what prices are "acceptable"; the second stage relates to the quality dimension—the consumer must decide what qualities are "acceptable." What makes the problem particularly interesting is that, in practice, these two dimensions—price and quality—tend to be related. This dependence means that the two stages of the consumer's problem are also related. Moreover, this dependence implies, as we shall show, that the optimal stopping rule at the first stage may well no longer be of the reservation-price form. For example, it may well be optimal for the consumer to search for expensive (rather than cheap) goods; alternatively, it could be optimal for the consumer to search for goods in some middle price range. Indeed, a variety of possible stopping rules may apply at the first stage; as we shall show, the form depends crucially on the form of the dependence between quality and price.

This two-stage problem belongs to a class of similar problems now beginning to appear in the literature. A similar consumer-search problem has been developed independently by Wilde (1977). In many respects his model is similar to ours, though it proceeds at a somewhat higher level of generality. Wilde's work follows from an earlier paper (Wilde 1979) on the parallel problem in the job-search area (a prob-

lem in which the wage is observed prior to accepting a job offer, and the nonpecuniary job characteristics are observed after experiencing the job). This latter problem was studied independently, and in a slightly different form, by McKenna (1978). Work has also been carried out in this area by Lippman and McCall (in press).

The paper is organized as follows: In the next section we introduce our notation and detail the assumptions of our model; the following section finds the solution to the consumer's two-stage problem and examines the properties of this solution; a concluding section reviews and appraises the key results.

II. The Model

We assume that the consumer-searcher is interested in two characteristics of the good—quality and price. We denote by Q the quality of the good as measured by the per period money value of the product's quality; we assume that the value of Q for a particular good can be observed only after purchase. The good yields Q for each period of its n-period lifetime; n is assumed to be the same for all goods. We denote by P the price of the good as measured by the price per unit per period of its lifetime (i.e., the per period "user cost" of the good).[1]

The consumer makes independent random selections from the joint distribution of P and Q; each drawing costs a (constant) amount c (> 0). The value of P is immediately revealed, while the value of Q is revealed only if purchase takes place. We assume that the searcher knows the joint distribution of P and Q over all goods: We denote by $g(\cdot)$ the marginal probability density function (pdf) of P, and by $f(\cdot \mid p)$ the conditional pdf of Q given p; we use $G(\cdot)$ and $F(\cdot \mid p)$ to denote the corresponding distribution functions and permit P and Q to take values on the positive real line.

The consumer faces an infinite horizon, discounts all values at the rate ρ ($0 < \rho < 1$), and wishes to have exactly one unit of the good in each period. Given these assumptions, the lifetime discounted value to the consumer is

$$\sum_{t=1}^{\infty} \rho^{t-1}(Q_t - P_t - S_t), \tag{1}$$

where Q_t and P_t are the quality and price of the good being consumed in period t, and S_t is the amount spent on search (possibly zero) in period t.

The individual's objective is to find, and follow, the strategy which

[1] Alternatively, the purchase price—payable at the time of purchase—is $\Sigma_{t=1}^{n} \rho^{t-1} P$ (where ρ, see below, is the discount rate).

maximizes the expected value of the expression (1). The model is built around the following assumed sequence of events:

1. The consumer searches, at a cost c per search, over values of P until an acceptable value of P is obtained. We assume that this search process is instantaneous and that no passage of time occurs.[2] Thus the consumer can obtain instantaneously n "quotes" at a total cost nc, whether the searching is done sequentially or not.
2. The consumer then purchases one unit of the good.
3. After purchase, the consumer experiences the good and thereby discovers the value of Q associated with the good. (We assume that all uncertainties are resolved after purchase.)
4. Having discovered Q, the consumer decides whether to keep the good for the whole of its n-period lifetime or to dispose of the good (at zero cost) at some point within its lifetime. If the latter option is used, the individual returns to stage 1.
5. If the good is kept until the end of its lifetime, the consumer must decide whether Q was sufficiently good for it to be repurchased. If so, the good is continually repurchased thereafter; otherwise, search continues again at stage 1.

The model is appropriate for regularly purchased nondurable goods for which the amount consumed per period is roughly constant, such as bread, cereals, eggs, washing powder, newspapers, and toilet paper. However, the purchase of some durable items such as televisions, cars, and washing machines may also be described in our framework.[3]

III. The Solution

The consumer's problem is a two-stage one: The first stage relates to the price dimension, the second stage to the quality dimension. Our solution similarly proceeds in two stages.

We begin with the first stage (that described by 1 in the sequence of events above), namely, the stage relating to the price dimension. The consumer's problem at this stage is to decide which price quotes are

[2] This is crucial in view of our earlier assumption that the consumer requires exactly one unit of the good per period. In contrast, Wilde (1977) postulates a time-consuming search process and assumes that the consumer derives no utility in periods when he does not have a unit of the good and derives utility $U(P, Q)$ in other periods.

[3] The repurchase of durable goods, however, may be made after a much longer period of consumption. For our model to be valid in describing purchase decisions of this type requires that we assume unchanging market conditions over long periods.

acceptable (in that purchase should take place at such quotes if obtained) and which price quotes are unacceptable (in that search should continue). In other words, the consumer must decide on the optimal acceptance set for prices. To find this optimal set, we first begin with an arbitrary acceptance set which we denote by $\mathcal{P}$ ($\subset \mathcal{R}^+$). This set $\mathcal{P}$ defines a stopping rule applicable at stage 1 as follows:

$$\begin{aligned} &\text{if } P \in \mathcal{P}, \text{ stop searching and purchase good at } P; \\ &\text{if } P \notin \mathcal{P}, \text{ continue searching.} \end{aligned} \tag{2}$$

Here P denotes the latest price quote obtained. We wish to find the best such set, which we denote by $\mathcal{P}^*$, in the sense that it maximizes the expected value of the objective function (1). To find this $\mathcal{P}^*$ we proceed as follows. Consider the consumer starting from a position in which no acceptable price has yet been obtained; suppose that the consumer uses the arbitrary acceptance set $\mathcal{P}$ with respect to the first purchase decision but thereafter behaves optimally with respect to all future decisions (including those relating to any future return to stage 1). Let $V(\mathcal{P})$ denote the expected value of the objective function following this sequence of events. Clearly $\mathcal{P}^*$, by definition, is a set for which $V(\mathcal{P})$ attains its maximum value; if we denote this latter by V^*, we have

$$\begin{aligned} &V^* \geq V(\mathcal{P}) \qquad \text{for all } \mathcal{P} \in 2^{\mathcal{R}^+}, \\ &V^* = V(\mathcal{P}^*). \end{aligned} \tag{3}$$

We will demonstrate that $\mathcal{P}^*$ need not be determined by a unique reservation price (i.e., $\mathcal{P}^*$ need not be of the form $\{P; P \leq P^*\}$). This is because the direct attraction of low prices may be more than offset by the indirect detraction of associated low qualities. In mathematical terminology, the return function from accepting P need not be monotonically decreasing in P because of the relationship between Q and P in the joint distribution. We pursue this formally below.

We now turn to the second stage (that described by 4 and 5 in the sequence of events above), namely, the stage relating to the quality dimension. The consumer's problem at this stage is to decide what qualities are acceptable, both in terms of keeping the good until the end of its lifetime and in terms of repurchasing that good at the end of its lifetime. In contrast with the decision at stage 1, these acceptance decisions do have the reservation property—the higher the Q, the better off the individual.[4]

[4] In contrast with the first stage where a higher P may make the individual better or worse off. For a formal demonstration of the textual assertion see Hey (1979) and/or Theorem 1 of Wilde (1979).

We wish to determine the optimal reservation qualities; to find *these we first begin with arbitrary reservation levels defined as follows. Let q_i denote an arbitrary reservation quality used at the end of period i following the purchase of the good ($i = 1, \ldots, n - 1$).*[5] *Using these arbitrary values, the decision at stage 4 is as follows:*

$$\begin{aligned} &\text{if } Q < q_i, \text{ dispose of the good at the end of period } i \text{ (if not already disposed of) and return to stage 1;} \\ &\text{if } Q \geq q_i, \text{ keep the good (at least until the end of period } i + 1). \end{aligned} \tag{4}$$

Here Q denotes the quality of the good purchased.

Further, let q_n denote an arbitrary reservation quality used at the end of period n (at the end of the life of the good). Using this arbitrary value, the decision at stage 5 is as follows:

$$\begin{aligned} &\text{if } Q < q_n, \text{ do not repurchase the good; instead return to stage 1;} \\ &\text{if } Q \geq q_n, \text{ continue repurchasing that good forever.} \end{aligned} \tag{5}$$

Although not essential, particularly for arbitrary q_i, it would seem appropriate to assume that $q_i \leq q_{i+1}$ ($i = 1, \ldots, n - 1$).

To find the consumer's optimal strategy, we need to find the best set of q_i ($i = 1, \ldots, n$), which we denote by q_i^* ($i = 1, \ldots, n$), in the sense that this set maximizes the expected value of the objective function (1). To find these q_i^* we proceed as follows. Consider the consumer starting from a position in which a purchase at price P has just been made. Suppose that the consumer uses arbitrary reservation qualities q_i ($i = 1, \ldots, n$) with respect to this purchase but thereafter behaves optimally with respect to all future decisions (including those relating to any future returns to stages 4 and 5). Let $W(P; q_1, \ldots, q_n)$ denote the expected value of the objective function following this sequence of events. Clearly, q_i^* ($i = 1, \ldots, n$) is a set for which $W(P; q_1, \ldots, q_n)$ attains its maximum value (for given P); if we denote this latter by $W^*(P)$, we have

$$\begin{aligned} &W^*(P) \geq W(P; q_1, \ldots, q_n) \quad \text{for all } q_i \in \mathscr{R}^+ \ (i = 1, \ldots, n), \\ &W^*(P) = W(P; q_1^*, \ldots, q_n^*). \end{aligned} \tag{6}$$

This completes our notational preliminaries, and we can now proceed to an investigation of $\mathscr{P}^*$ and the q_i^*.

If the stopping set $\mathscr{P}$ is used at stage 1, we have

$$V(\mathscr{P}) = -c + V(\mathscr{P})[1 - \textstyle\int_{\mathscr{P}} g(p)dp] + \int_{\mathscr{P}} W^*(p)g(p)dp. \tag{7}$$

[5] We exclude $i = 0$ on grounds of realism; its inclusion would not materially affect any of the results.

This equation states that, starting from a position in which no acceptable price quote has been found, the individual searches once more at cost c and either receives an unacceptable quote (returning to "square one") or receives an acceptable quote with expected value of $W^*(p)$.[6] Equation (7) gives:

$$V(\mathcal{P}) = [-c + \int_{\mathcal{P}} W^*(p)g(p)dp]/\int_{\mathcal{P}} g(p)dp. \tag{8}$$

Now, it can easily be shown (Hey 1979) that an optimal set is given by

$$\mathcal{P}^* = \{P;\ W^*(P) \geq V^*\}, \text{ where } c = \int_{\mathcal{P}^*} [W^*(p) - V^*]g(p)dp. \tag{9}$$

The first of these simply states that a price quote should be accepted if the expected value (of the objective function) obtainable by accepting it is at least as large as that obtainable by rejecting it. The value of V^* is determined by (9); we assume that this is positive (otherwise it is not optimal for the consumer to initiate search for the good).

Examination of (9) shows that the form of the set $\mathcal{P}^*$ depends crucially on the behavior of the function W^*. The familiar case is when W^* is a monotonically decreasing function (i.e., when higher prices unambiguously make the consumer worse off); in this case, as can be seen from (9), the set $\mathcal{P}^*$ is given by $\{P;\ P \leq P^*\}$, where P^* is the familiar reservation price. However, in our model there is nothing to guarantee that W^* is monotonically decreasing; indeed, W^* may take a variety of different forms, depending on the relationship between quality and price in the joint distribution. In the extreme, it is possible (although unlikely) that quality increases so fast with price that the function W^* is monotonically increasing; in such a case the acceptance set $\mathcal{P}^*$ would take the form $\{P;\ P \geq P^*\}$ where P^* is a lower reservation price—a buy-expensive policy would be optimal. The general case (one in which W^* has both increasing and decreasing segments) is illustrated in figure 1.

As the form of the function W^* is crucial to the consumer's optimal behavior, we now proceed to an investigation of its properties. From our story of consumer behavior, we have

$$\begin{aligned} W(p;\ q_1, \ldots, q_n) = {} & \left[-p(1-\rho^n)/(1-\rho) + \int_0^x qf(q \mid p)dq \right] \\ & + \Sigma_{i=1}^{n-1} \rho^i \left[V^* \int_{q_{i-1}}^{q_i} f(q \mid p)dq + \int_{q_i}^x qf(q \mid p)dq \right] \\ & + \rho^n V^* \int_{q_{n-1}}^{q_n} f(q \mid p)dq + \rho^n \int_{q_n}^x [(q-p)/(1-\rho)]f(q \mid p)dq \end{aligned} \tag{10}$$

(where $q_0 \equiv 0$).

[6] Note that the search process is assumed to be instantaneous and thus the expected values following the search are not discounted; see also n 2 above. Note further that if $\mathcal{P} = \phi$, then (7) yields $V(\phi) = -c + V(\phi)$, which (if $c > 0$) implies that $V(\phi)$ is minus infinity.

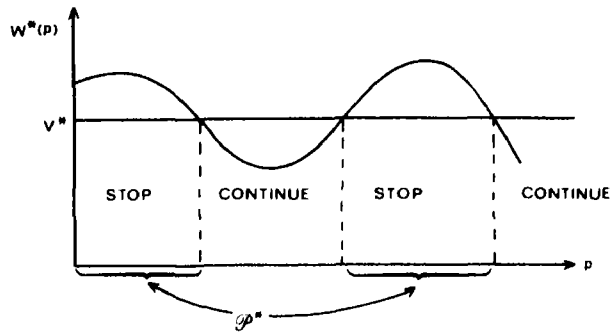


FIG. 1.—Derivation of the set $\mathcal{P}^*$ in the general case

The optimal reservation qualities at the end of periods 1 to $(n - 1)$ —the q_i^* $(i = 1, \ldots, n - 1)$—are found by setting $\partial W/\partial q_i = 0$. It is easy to show that this implies that

$$q_i^* = V^*(1 - \rho) \qquad (i = 1, \ldots, n - 1). \tag{11}$$

Two things of interest flow from this: First, $q_1^* = \ldots = q_{n-1}^*$, which means that if the good's quality is sufficiently high for it not to be thrown away at the end of the first period of its lifetime, it should be kept for the remainder of its lifetime. (This result is a consequence of our assumed constancy of the per period quality rate; if, in contrast, quality were to decline as the good got older, this strong result would no longer hold.) Second, $q_1^* = V^*(1 - \rho)$, which means that a good should be thrown away (before the end of its useful life) if and only if its quality is lower than the per period value obtainable from starting search afresh.

The optimal reservation quality at the end of the good's lifetime is found by setting $\partial W/\partial q_n = 0$; this implies that

$$q_n^* = V^*(1 - \rho) + p, \tag{12}$$

which states that the consumer should *re*purchase the good if and only if its quality net of price is at least as large as the expected value obtainable from starting search afresh. Both this, and the earlier result (11), accord with intuition.

If we now substitute (11) and (12) into (10), we get an expression for W^*; after some simplification, this yields

$$\begin{aligned} W^*(p) = \rho V^* &- [p(1 - \rho^n)/(1 - \rho)] + \int_0^\infty [1 - F(q \mid p)]dq \\ &+ [\rho(1 - \rho^{n-1})/(1 - \rho)] \int_{q_1^*}^\infty [1 - F(q \mid p)]dq \\ &+ [\rho^n/(1 - \rho)] \int_{q_n^*}^\infty [1 - F(q \mid p)]dq \end{aligned} \tag{13}$$

(where, in view of the equality of the q_i^* for $i = 1, \ldots, n-1$, we simply work with q_1^* and q_n^*).

Equations (9), (11), (12), and (13) solve for V^*, $W^*(\cdot)$, q_1^*, q_n^*, and $\mathcal{P}^*$ in terms of c, ρ, n, $f(\cdot \mid \cdot)$, and $g(\cdot)$; they provide the complete solution to our model.

The form of (13) can now be investigated; from it and (12) we have

$$\begin{aligned} W^{*\prime}(p) = & -(1-\rho^n)/(1-\rho) - \int_0^\infty \frac{\partial F(q \mid p)}{\partial p} dq \\ & - [\rho(1-\rho^{n-1})/(1-\rho)] \int_{q_1^*}^\infty \frac{\partial F(q \mid p)}{\partial p} dq \\ & - [\rho^n/(1-\rho)] \int_{q_n^*}^\infty \frac{\partial F(q \mid p)}{\partial p} dq \\ & - [\rho^n/(1-\rho)][1 - F(q_n^* \mid p)]. \end{aligned} \tag{14}$$

Inspection of this shows that $W^{*\prime}$ is unambiguously negative if $\partial F(q \mid p)/\partial p$ is positive everywhere; however, this latter condition implies that $F(\cdot \mid p)$ is stochastically decreasing in p—or, in economic terms, that quality is stochastically decreasing in price. This is highly unlikely in practice. A more realistic case is to assume that quality generally increases with price so that $F(\cdot \mid p)$ is stochastically increasing in p (i.e., $F(t \mid p_1) \geq F(t \mid p_2)$ for all t whenever $p_1 < p_2$; see Hadar and Russell [1969]). However, this restriction is not sufficient to sign $W^{*\prime}$. On reflection, this is not a surprising result; one can imagine cases where quality increases very strongly with price (so that $W^{*\prime}$ is positive) and, also, cases where quality increases very slowly with price (so that $W^{*\prime}$ is negative—the direct price effect outweighing the indirect quality effect). In general, it is possible that $W^{*\prime}$ has both positive and negative sections, in which case the stopping rule at stage 1 would be of a form similar to that shown in figure 1.

Wilde (1977, 1979) also discusses similar possibilities (see, in particular, Wilde [1979], n. 11) Moreover, because his model proceeds at a rather higher level of generality than our own (namely, his objective function is not simply linear in quality and price), he finds restrictions (sufficient to sign his equivalent of our $W^{*\prime}$) on the joint distribution virtually impossible to obtain. Instead, he simply assumes that the distribution is such that the return function (W^*) is monotonically decreasing.

However, some further insight can be gained as follows.[7] Denote by Q_p the (random) quality associated with the price p and denote the

[7] Although our model assumes that "utility" is linear in price and quality, our results would still go through if utility was a separable function of price and quality if the distributions $f(\cdot \mid \cdot)$ and $g(\cdot)$ were reinterpreted appropriately.

distribution function of Q_p by $F_p(\cdot) \equiv F(\cdot \mid p)$. Suppose that the conditional distributions are such that

$$F_p(q) = F_0(q - kp), \tag{15}$$

that is, Q_p is a translation of Q_0. (Thus the distribution of Q_p is the same as that of Q_0 shifted to the right by a distance kp.) It is a simple matter to show that

$$\partial F(q \mid p)/\partial p = -kf(q \mid p). \tag{16}$$

If this is substituted into (14), we find, after some simplification, that

$$\begin{aligned} W^{*\prime}(p) = (k - 1)[1 - \rho^n F(q_n^* \mid p)]/(1 - \rho) \\ - k\rho(1 - \rho^{n-1})F(q_1^* \mid p)/(1 - \rho). \end{aligned} \tag{17}$$

There is an important special case in which the second term in (17) disappears; this is when $F(q_1^* \mid p) \equiv 0$, or, in economic terms, when there is no possibility of finding a quality so bad that it pays to throw the good away before the end of its lifetime.[8] In such a case, equation (17) implies that

$$W^{*\prime}(p) \gtreqless 0 \text{ according as } k \gtreqless 1. \tag{18}$$

This result makes good sense: It states that the return function is an increasing (decreasing) function of price if $q - p$ (the quality net of price) is an increasing (decreasing) function of price. Moreover, if $q - p$ is independent of price (that is, if $k = 1$), W^* is horizontal (at the level V^*), and the consumer simply accepts the first price quote received whatever its value.

The general case is when $F(q_1^* \mid p) > 0$, that is, when there exists a positive probability of getting a good whose quality is so low that it pays to throw it away and start search afresh. In this general case, the simple result (18) no longer holds. However, as inspection of (17) shows, if k is less than or equal to one, then $W^{*\prime}$ is strictly negative. Thus, if "net quality" $(q - p)$ decreases with price, the return function is monotonically decreasing and a buy-cheap policy remains optimal. However, k larger than 1 is not now sufficient for $W^{*\prime}$ to be positive, because the (nonzero) possibility of getting a "lemon" destroys the generally greater attractiveness of higher-price goods.[9] However, by differentiating (17) (and using [16]) it can be shown that k greater than 1 is a sufficient condition for $W^{*\prime\prime}$ to be positive. Furthermore, from (17) it can be seen that $\lim_{p\to\infty} W^{*\prime}(p) = (k - 1)/(1 - \rho) > 0$ if $k >$

[8] This effectively is the case considered by Wilde (1977).

[9] A sufficient condition for $W^{*\prime}$ to be positive is that k is at least as large as $(1 - \rho^n)/(1 - \rho)$, so that net quality increases sufficiently fast with price to outweigh the nonzero lemon possibility.

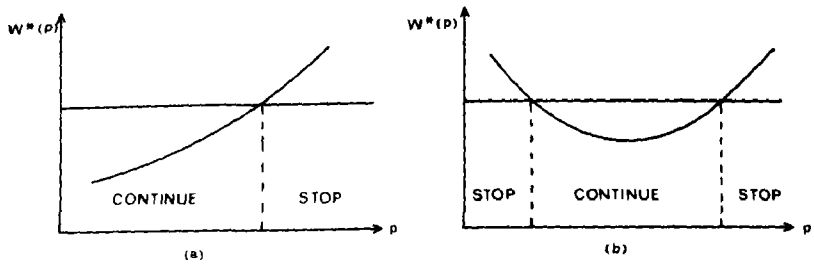


FIG. 2.—Derivation of $\mathcal{P}^*$ when (15) holds with $k > 1$

1.[10] These results imply that if k is greater than 1 (i.e., if net quality increases with price), then $W^{*\prime}$ is monotonically increasing and that even if it is negative at lower values of p, it must become positive at some higher value. Thus, if k is greater than 1, one of the two possibilities pictured in figure 2 must exist.

Case b in figure 2 is particularly interesting, implying that the consumer should search until either a very low, or a very high, price quote is obtained. Goods at intermediate prices should not be purchased. This result follows from the combination of net quality increasing with price and the lemon possibility: High prices are attractive because the associated high qualities are a sufficiently great incentive to outweigh the possible lemon effect; low prices are attractive despite the associated low qualities because of the small loss involved in throwing away a lemon which cost little in the first place; intermediate prices are not attractive because they fall between these two stools.

The results discussed in the paragraphs above are, of course, specific to the conditional distribution formulation given by (15). However, these results do yield some insight into the general considerations underlying the form of the function W^*; clearly important is the (stochastic) relationship between net quality $(q - p)$ and price. Alternative formulations can easily be envisaged. For example, imagine that the conditional distribution of net quality increases at an increasing rate with price up to some intermediate price and thereafter increases at a decreasing rate. In such a case, W^* may be an increasing function for low p and a decreasing function for high p; the price-acceptance set would then consist of intermediate prices, both very low and very high prices being rejected in favor of further search.

Clearly, a whole variety of possibilities exist: Depending on the

[10] Since $F_p(q)$, as defined in (15), approaches zero as p approaches infinity for any finite q.

conditional distribution of quality given price, buy cheap, buy expensive, buy medium, or various combinations of these are all plausible buying strategies. Thus "gaps" in the stopping set $\mathcal{P}^*$ can be quite consistent with optimal behavior.

Before concluding we note the following comparative static results. They can be derived from equations (9), (11), (12), and (13); the detailed proofs can be found in Hey and McKenna (1979).

1. Each of the following makes the consumer-searcher better off in that V^* and $W^*(p)$ increase: (i) a decrease in the search cost c, (ii) a decrease in the lifetime of the good n, (iii) a rightward shift of the conditional distributions of Q_p, (iv) a rightward shift of the marginal distributions of P if W^* is a monotonically increasing function, (v) a leftward shift of the marginal distribution of P if W^* is a monotonically decreasing function, (vi) an increase in risk (in the Rothschild and Stiglitz [1970] sense) in the conditional distributions of Q_p. Consequently the consumer becomes more choosy in that q_1^* and q_n^* increase and the set $\mathcal{P}^*$ gets smaller.

2. The expressions $V^*(1 - \rho)$ and $W^*(p)(1 - \rho)$ increase if ρ increases, that is, if the consumer places greater weight on the future; consequently, the consumer gets more choosy in the present in that q_1^* and q_n^* increase and the set $\mathcal{P}^*$ gets smaller.

All of these results accord with intuition, with the possible exception of that relating to n. However, when it is remembered that P is the price per period of the good's lifetime, it does make sense to conclude that a smaller n is better, in that it reduces the time spent "locked in" to an unsatisfactory good.

IV. Summary and Conclusions

This paper has presented a unified model of consumer search with both price and quality uncertainty. Because price is revealed prior to purchase while quality is not revealed until after purchase, the consumer's problem is essentially a two-stage problem. The first stage relates to the price and involves choosing an acceptance set for prices. The second stage relates to the quality, and involves deciding whether a particular quality is sufficiently good (*a*) to retain it after purchase and (*b*) to continue buying it thereafter. The problem is made interesting and nontrivial by the dependence of quality on price.

The major result is the finding that the price-acceptance set is not necessarily connected and that the conventional reservation price rule is optimal only in some circumstances. In the general case, depending on the stochastic relationship between quality and price, buy cheap, buy expensive, buy medium, or various combinations of these are all plausible buying strategies.

In the form presented in the paper, the consumer's utility is assumed to be linear in price and quality. However, all our results still hold if utility is a separable function of price and quality, as long as the distributions of price and quality are reinterpreted appropriately (as distributions of the utility of price and the utility of quality). Of course, in such a generalization optimal behavior would vary from consumer to consumer, depending on the differences between their utility functions. It would be of interest to explore the implications for price/quality equilibria in markets resulting from such behavior. It is to be hoped that this paper has laid a firm foundation for such an exploration.

References

Hadar, Josef, and Russell, William R. "Rules for Ordering Uncertain Prospects." *A.E.R.* 59 (March 1969): 25–34.

Hey, John D. "A Simple Generalised Stopping Rule." *Econ. Letters* 2 (July 1979): 115–20.

Hey, John D., and McKenna, Chris J. "Consumer Search with Uncertain Product Quality." Discussion paper no. 33, Univ. York, Inst. Soc. and Econ. Res./Dept. Econ., 1979.

Lippman, Steven A., and McCall, John J. "The Economics of Belated Information." *Internat. Econ. Rev.* (in press).

McKenna, Chris J. "A Theory of Contractual Decision under Uncertainty." Discussion paper no. 27, Univ. York, Inst. Soc. and Econ. Res./Dept. Econ., 1978.

Rothschild, Michael, and Stiglitz, Joseph E. "Increasing Risk. I: A Definition." *J. Econ. Theory* 2 (September 1970): 225–43.

Wilde, Louis L. "On the Formal Theory of Inspection and Evaluation in Product Markets." Mimeographed. Pasadena: California Inst. Tech., 1977.

——— "An Information-theoretic Approach to Job Quits." In *Studies in the Economics of Search*, edited by Steven A. Lippman and John J. McCall. Amsterdam: North-Holland, 1979.

Commodity-Choice Behavior with Pigeons as Subjects

Raymond C. Battalio and John H. Kagel

Texas A&M University

Howard Rachlin

State University of New York at Stony Brook

Leonard Green

Washington University

Starting from an initial (baseline) budget line, income-compensated price changes always resulted in substitution effects consistent with the Slutsky-Hicks theory. This behavior cannot be explained by a simple random-behavior model. Similar changes in relative prices that did not originate from the initial (baseline) budget line resulted in "undersubstitution effects": The composition of consumption changed in the expected direction, but the magnitude of change was not large enough to be consistent with the initial commodity bundle chosen. These undersubstitution effects are not explainable by shifting preference patterns or anchoring effects found in inconsistent choice sequences with human subjects.

This paper reports tests of the negativity of the Slutsky-substitution effect in commodity-choice experiments with pigeons as subjects. Starting from a given (baseline) budget line, income-compensated

Research support from the National Science Foundation is gratefully acknowledged. An earlier version of this paper was presented at the 1979 Summer Econometric Society meetings in Montreal. We thank the editor and the referee for helpful comments. The usual caveat applies.

[*Journal of Political Economy*, 1981, vol. 89, no. 1]
© 1981 by The University of Chicago 0022-3808/81/8901-0001$01.50

increases or decreases in relative commodity prices always resulted in substitution effects consistent with the Slutsky-Hicks theory. These results, which replicate earlier studies using rats as subjects (Kagel et al. 1975, 1980), are not explainable by simple random-behavior models proposed in the literature (Becker 1962; Chant 1963).

Similar changes in relative prices that did not originate from the baseline resulted in "undersubstitution effects"; that is, the composition of consumption tended to change in the expected direction, but the changes were not large enough to be consistent with the original (baseline) commodity bundle consumed. These undersubstitution effects are not explainable in terms of shifting preference patterns or anchoring effects (Tversky and Kahneman 1974) found in inconsistent choice sequences with humans as subjects. Further experiments will be required to identify the basis for this behavior.

Data on learning patterns as subjects adjust to changes in experimental conditions are also reported. Questions of interest here concern identifying uniformities in both the speed of the adjustment process and its pattern, whether it can be characterized by concepts of habit formation or stock-adjustment processes (Houthakker and Taylor 1970). The nature of the learning process is, of course, strictly an empirical matter in terms of the static Slutsky-Hicks theory.

Space considerations do not permit a detailed discussion of the rationale behind using nonhuman subjects to investigate economic theories (see Kagel and Battalio [1980] for this argument). For the more skeptical reader we note that if one defines economics as ". . . the study of the allocation of scarce resources among unlimited and competing uses" (Rees 1968), then animal psychologists, ecologists, and biologists have been involved in studying economic behavior for some time now (Rapport and Turner 1977; Hirshleifer 1978; Rachlin 1980). It is but a small step to take the technologies of these related disciplines and apply them to behavior of interest to economists, for example, commodity-choice behavior. At a minimum, such studies expand considerably the scope for comparative economic analysis. At a maximum they provide a laboratory for identifying, testing, and better understanding general laws of economic behavior. Use of this laboratory is predicated on the fact that behavior as well as structure vary continuously across species, and that principles of economic behavior would be unique among behavioral principles if they did not apply, with some variation, of course, to the behavior of nonhumans.

The plan of the paper is as follows: Section I describes the hypotheses tested and the procedures used to operationalize the budget constraint. Sections II and III report the two sets of experiments conducted: Both comparative static and temporal patterns of behavior are examined. We end with a discussion which summarizes the test outcomes and their implications for future research and returns to

the question of the relevance of the experiments for understanding human economic activity.

I. Hypotheses Tested and Experimental Procedures Employed

A necessary condition that all members of the class of Slutsky-Hicks demand functions must satisfy is that for income-compensated price changes (of the Slutsky type) the consumer purchase a commodity bundle that was previously unattainable. More precisely,

$$\text{if } p^i \neq p^j \text{ and } p^j x^i = p^j x^j \text{ then } (p^j - p^i)(x^j - x^i) < 0, \tag{1}$$

where p^i and x^i represent vectors of prices and quantities of goods purchased in period i. Condition (1), also referred to as the weak axiom of revealed preference, provides the basis for a direct consistency test of the theory (Hicks 1956).

Condition (1) was tested by having food- and water-deprived pigeons choose between the consumption of these two commodities. The subjects were placed in an experimental chamber for a fixed period of time each day. Each chamber contained a food hopper and water reservoir located on the front panel. Access to the food and water was controlled by separate variable time schedules. Under a variable time schedule commodities become available for a fixed period of time every α seconds on the average, where α is determined as part of the experimental design. Only one of the variable time schedules was in operation at a given time, and subjects could switch between the food and water schedule by a single peck on a control key. Pecks on this key were not required to continue food or water deliveries, but only to change which commodity would be delivered. Color-coded lights were used to signal whether the food or water schedule was in effect. The price of each commodity was varied by altering the average time between deliveries of that commodity (access times remaining constant), while the income constraint was the total time available for the delivery of the two goods.

Under these procedures time spent under the food and water schedules (T_F, T_W) produces commodities at an average rate of

$$C_F = \frac{1}{\alpha_F} T_F, \qquad C_W = \frac{1}{\alpha_W} T_W, \tag{2}$$

where α_F, α_W are the average values of the variable time schedules, and C_F and C_W are food and water consumption measured in number of payoffs. Since time required to switch between schedules is sufficiently small that it can be ignored for all practical purposes, substituting (2) into the total time (T_T) constraint

$$T_T = T_W + T_F \tag{3}$$

gives an expected value for the budget constraint of

$$\alpha_F C_F + \alpha_W C_W = T_T. \tag{4}$$

From (4) it is clear that increasing (decreasing) the average time interval between deliveries of a commodity has the same impact as a price increase (decrease) in the usual specification of the budget constraint. Analogously, increasing (decreasing) the total time available for delivery of the goods increases (decreases) nominal income (see Becker [1971] for the comparative statics of this problem).

To establish conditions for testing the negativity of the Slutsky-substitution effect, consumption patterns were first recorded under a set of initial (baseline) price (α_F^i, α_W^i) and income (T_T^i) conditions. The price of one of the commodities was then increased, accompanied by a corresponding decrease in the price of the other commodity so as to achieve, as nearly as possible, perfect compensation

$$\alpha_F^i C_F^i + \alpha_W^i C_W^i = \alpha_F^j C_F^j + \alpha_W^j C_W^j, \tag{5}$$

where the superscripts refer to the initial (i) and price-change period (j). Since commodity consumption varied somewhat from day to day under constant experimental conditions, the median bundle consumed during the last 5 days of a condition served as the reference point for compensation. Total time available for consumption remained constant across experimental conditions, as did the amounts of food and water delivered per payoff.

Consistent choice behavior (a negative Slutsky-substitution effect) requires that

$$T_T^i - (\alpha_F^i C_F^j + \alpha_W^i C_W^j) < 0. \tag{6}$$

That is, with the compensated change in relative prices the consumer purchases a commodity bundle in the price-change period that it could not afford to purchase during baseline. Since, as already noted, consumption was not the same from day to day under constant experimental conditions, t-tests applied to the mean value of (6) calculated over the last 10 days of the price-change period were used to determine consistency.

In cases where (6) is negative, a simple random-behavior process could explain the results (Becker 1962; Chant 1963). For example, if the pigeons randomly choose the amount of time spent receiving each of the commodities, and total time remains constant across experimental conditions, then reducing the time between food deliveries, which corresponds to a reduction of food costs, will result in increased food consumption. Correspondingly, increasing the time between water payoffs, which increases its price, would result in a decrease in average water consumption. Under random behavior the

expected value of time spent receiving each of the commodities would be

$$\frac{\alpha_k C_k}{T_T} = q_k \qquad \begin{matrix} k = F, W \\ 0 \leq q_k \leq 1 \end{matrix}, \tag{7}$$

where q_k is constant across experimental conditions.[1]

Chant has characterized consumers who behave in this fashion as "money deciders." In the case where $q_k = .5$, the pigeons would be spending an equal amount of time collecting each of the commodities. In Chant's terminology they would be "impulsive money deciders." For the more general case, where the pigeons spend a constant but unequal proportion of time collecting each of the commodities, they would be characterized as "proportionately inert money deciders." In cases where (6) is negative, we will test whether behavior is significantly different from (7) using t-tests for paired comparisons and appropriate analysis-of-variance procedures for comparing across several conditions. Rejection of (7) suggests that something other than a simple random behavioral process is guiding choices.

II. Experiment 1

Procedures

Four male White Carneaux pigeons with no previous experimental history served as subjects (S's). Subjects were completely dependent on food and water consumed during experimental sessions with no access to these or any other commodities in their home cages. Experimental sessions were run once a day 7 days a week starting at approximately the same time each day. Consequently, subjects' body weight varied somewhat between conditions depending on consumption patterns.

All S's had extensive training with the procedures prior to the start of the experiment. Experimental conditions were changed when inspection of graphs of the data indicated an absence of any significant drift in consumption patterns, although at times conditions were maintained for considerably longer periods simply to see what would happen.

Food payoffs consisted of 8 seconds access to a food hopper containing mixed pigeon grains. Fluid payoffs consisted of 3 seconds access to a reservoir containing tap water. Access to the food hopper

[1] Negative values for (6) rule out the cases of "inert" or "impulsive goods deciders" (Becker 1962; Chant 1963). Of course, behavior in accordance with (7) is observationally indistinguishable from behavior being motivated by a Cobb-Douglas utility function (Becker 1962).

and water reservoir was blocked at all other times. Time spent eating and drinking did not count against total time available for delivery of commodities (T_T), which was 60 minutes for all S's.[2]

For three of the four S's, food and water were delivered on the average at 60-second intervals during the initial (baseline) period. This was followed by an increase in the relative price of food. Baseline conditions were then reinstated, followed by a decrease in relative food prices, which was followed by a return to baseline conditions. (Table 1 shows the sequence and magnitude of relative price changes.) To account for any shifts in preference structure, the median commodity bundle taken in the last 5 days of each baseline served as the reference point for the Slutsky-compensated price change of the following period. A fourth subject, S 85, started with food payoffs coming twice as often as water deliveries (relative price of food equal to .5), followed by a Slutsky-compensated increase in the price of food. Baseline conditions faced by the other birds were implemented in period 3 for this subject. Since the mean commodity bundle purchased during period 3 could very nearly have been bought during the previous period, we compare these two points also.

Comparative Static Results

The effects of the Slutsky-compensated price changes are reported in table 1. The results of the increase in food price relative to baseline conditions are reported between conditions 1 and 2 for S's 14, 46, and 50 and between conditions 2 and 3 for S 85. In all cases the weak-axiom test statistic (6) has the correct sign and is statistically significant at the .05 level for three of the four S's. Price elasticities for water ranged from a low of −.07 for S 50 to a high of −.81 for S 14, averaging −.38 across all four birds. These price elasticities are a bit larger than the highest values reported for food-fluid experiments with rats (Kagel et al. 1980). The increase in the relative price of food for S 85 between periods 1 and 2 also resulted in substitution effects in the expected direction.

Decreases in the relative price of food for S's 14, 46, and 50 in period 4 resulted in a negative (correct) sign for the weak-axiom test statistic, which was significant at the .05 level or better in all cases. Price elasticities here ranged from a low of −.10 for S 50 to a high of −.24 for S 14, averaging −.16 across S's.

Replication of baseline conditions following the price-change

[2] Initial values were determined so that S's could maintain healthy body weights from food and water consumed within the sessions. Values were determined on the basis of past experiments and experience during the training periods. Additional procedural details may be found in the Appendix.

TABLE 1

Results of Consistency Tests: Experiment 1

Subject and Experimental Condition	α_F/α_W	Weak-Axiom Test (t-Statistic)[a]	Mean Payoffs[b] (S_M) Food	Mean Payoffs[b] (S_M) Water	F-Test for Equality of Baselines (Marginal Significance Level)
Subject 14.					
1	1 0	−636 (2.44)**	46.4 (1.54)	13.6 (1.54)	3.36** (.05)
2	2.88		39 5 (2.20)	31.1 (6.55)	
3	1.0	−312 (3 26)***	37 1 (3.05)	23 3 (2.92)	
4	28		48.7 (2 21)	16.5 (82)	
5	1.0		42 4 (1 31)	18.2 (1.16)	
Subject 46					
1	1.0	−498 (2.74)**	34 2 (1.47)	25.8 (1 47)	94 (.40)
2	2.86		32.8 (1.63)	35.5 (4.65)	
3	1.0	−288 (2 79)**	37 2 (1.58)	22.9 (1 57)	
4	.22		43.7 (2.08)	21 1 (.41)	
5	1.0		36 0 (1.61)	23 4 (1 64)	
Subject 50:					
1	1.0	−48 (30)	41.1 (1.62)	19 1 (1.55)	2.39 (.11)
2	4.0		40.0 (98)	20 8 (3 66)	
3	1.0	−168 (2 43)**	42.6 (.82)	17 7 (.92)	
4	.32		47.2 (1.58)	15.6 (.50)	
5	1 0		44 7 (1.04)	16 0 (.94)	
Subject 85:					
1	5	−551 (1.87)*	33.5 (1.59)	33 4 (.82)	N.A
2	4.5		31 3 (1.33)	42.0 (4.74)	
3	1.0	−798 (3.88)***	31.4 (3 28)	28.2 (3.31)	

Note: α_F/α_W = relative food cost, N A = not applicable, S_M = standard error of the mean.
[a] The t-tests are two tailed with 9 degrees of freedom
[b] Calculated over the last 10 days of each period
* Significant at the 10 percent level.
** Significant at the 5 percent level
*** Significant at the 1 percent level

TABLE 2

TEST OF SIMPLE RANDOM-BEHAVIOR MODEL

Subject and Grouping*	Income Spent on Food (%)	Experimental Condition	Relative Price of Food	F-Test for Equal Percentage Hypothesis (Marginal Significance Level)
Subject 14				
A	.79	2	2.88	
A	.77	1	1.00	
A, B	.71	5	1.00	13.4†
B	.62	3	1.00	(.001)
C	.46	4	.28	
Subject 46				
A	.73	2	2.86	
B	.62	3	1.00	
B	.60	5	1.00	32.7
B	.57	1	1.00	(.001)
C	.32	4	.22	
Subject 50				
A	.89	2	4.00	
B	.75	5	1.00	
B	.71	3	1.00	55.4
B	.69	1	1.00	(.001)
C	.50	4	.32	
Subject 85				
A	.78	2	4.50	
B	.52	3	1.00	34.8
C	.34	1	.50	(.001)

* Conditions with the same grouping letter have the same percent of total expenditure devoted to food according to Duncan's multiple-range test ($\alpha = .05$)

† F-statistic excluding period 1 equals 14.7

periods showed consumption patterns to be reversible, at least directionally, in all cases. Preferences were sufficiently stable that for S's 46 and 50 there were no significant differences in consumption patterns between baseline periods. While consumption for S 14 in period 1 is found to differ significantly from that of periods 3 and 5 (at the .05 level using Duncan's multiple-range test [Winer 1971]), there are no differences between periods 3 and 5. We note that the ratio of food to water consumption for S 14 in period 1 was the highest value found across all S's in both experiments under baseline prices.

Table 2 reports tests of the simple random-behavior model (7). In all cases F-tests show that we can decisively reject the hypothesis that the percentage of total income spent on commodities was the same across differing experimental conditions. Using the letters A, B, and C to indicate experimental conditions that can be grouped together as having the same percentage of income spent on food (at the .05 significance level according to a Duncan multiple-range test), we see that the price-change periods are responsible for the differences in

each case. For S's 46 and 50 the percent of total income spent on food remained constant across baseline periods (conditions 1, 3, and 5) but increased significantly in period 2 when relative food prices increased and dropped significantly in period 4 when food prices dropped. Subjects 14 and 85 show similar response patterns with the exception, already noted, that consumption patterns differed between baseline periods for S 14. In all cases demand is far too inelastic for consistency to be attributed to the fact that S's spent a constant proportion of their time collecting each of the commodities. These pigeons cannot be characterized as either "impulsive" or "proportionately inert" money deciders.

Adjustment Patterns

The comparative static analysis tells us nothing about the learning path as subjects adjust to changes in relative prices. To investigate this we fit, within a given experimental condition, the following hyperbolic function to the commodity whose relative price decreased:

$$C_{kt} = a + b(1/t),$$

$$C_{kt} = \text{consumption of commodity } k \text{ on day } t\text{, commodity } k \text{ being the good whose relative price decreased,} \tag{8}$$

$$t = \text{time in days within a given experimental condition.}$$

The function (8) builds in the assumption that consumption approaches a steady-state value as time passes and there are no further changes in experimental conditions. We fit these functions because (1) in terms of conducting the experiments, conditions were only changed when inspection of the data suggested no significant drift in consumption patterns, and (2) graphs of the data suggested that in most cases we did not need to use more complicated functions to adequately represent the time path.

The coefficient b of (8) is readily interpretable in terms of a number of concepts within the economics and psychology literature. If b is not significantly different from zero we have, in effect, instantaneous, or at least very rapid, adjustment to the price changes, relative to the day-to-day variability in consumption patterns. Fitting (8) to the commodity whose relative price decreased under a given experimental condition and finding a negative value for b would imply that subjects responded gradually to the change in relative prices, increasing consumption of the now cheaper good with the passage of time. Such behavior would be consistent with concepts of habit formation found useful in describing human consumption for a number of nondurable goods (Houthakker and Taylor 1970; Phlips 1974).

The presence of habit formation would also be consistent with psychological notions that it takes S's time to perceive the changes in relative prices and even more time to learn the implications of responding to these changes.

In contrast, a positive sign for b in (8) would indicate that the immediate response to a decrease in relative prices was to consume more of the commodity than with the passage of time. This sort of response pattern goes under the heading of a stock-adjustment effect within the Houthakker-Taylor framework. It might be present in our data for any number of reasons.

In general there is no a priori basis for deciding whether habit-formation or stock-adjustment effects will predominate for a particular commodity. For example, Houthakker and Taylor (1970) found that housing exhibited strong habit-formation effects, while domestic services and water use showed stock-adjustment effects generally considered to be characteristic of consumer durable goods. The empirical work in psychology shows much the same diversity since learning curves of almost every conceivable shape have been reported in the literature (Mazur and Hastie 1978).

Figure 1 shows the daily water consumption for condition 2 when the relative price of water decreased. The mean consumption of water during the last 10 days of the previous period is also shown as an open circle on the water axis. The estimated b values and the Durbin-Watson statistics obtained from the least-squares estimates of equation (8) are also given for each subject.[3] The data show little trend and a great deal of variability. This is evident in the least-squares estimates of the time-trend coefficients, b, which are not statistically significant for any of the S's. Note, however, that in each case b is positive in sign, so that what little trend there is in the data is inconsistent with the notion that habit formation dominates choices. That is, there is a tendency to have more substitution into water during the first few days than is shown in the comparative static comparisons.

Figure 2 shows the daily food consumption for condition 4 when the relative price of food decreased. Once again there appears to be little trend, a result captured in the t-test ratios for the least-squares estimates of b, which are insignificant in all cases, although the signs of the coefficients are somewhat more favorable to the habit-formation hypothesis.

The absence of significant time-trend coefficients, in conjunction with the data reported in table 1, implies that S's adjusted very quickly to the changes in relative prices, at least within the time frame for

[3] The breaks in the sequential data plotted in figs. 1–4 correspond to days when the equipment malfunctioned. Dummy variables are used in these cases when adjusting (8) to the data

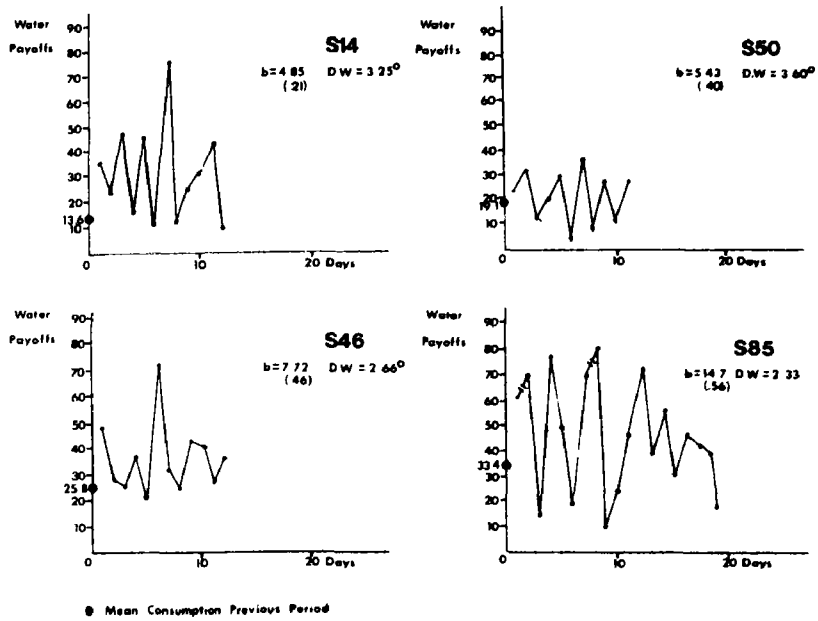


FIG. 1.—Temporal pattern of water consumption following a decrease in its relative price: experiment 1, condition 2. Broken lines indicate days with equipment failures. Estimated values of the time trend coefficient, *b*, with the corresponding *t*-statistic reported in brackets are given for each subject along with the value of the Durbin-Watson statistic. The symbols +, *, and ** indicate statistical significance at the 10, 5, and 1 percent levels, respectively. The symbols – and o indicate the Durbin-Watson test was indeterminate or there was an insufficient number of observations for the test.

which we have data (daily totals) and relative to the day-to-day variability in the data. For example, in going from condition 1 to 2, *S* 14 more than doubled water intake but has a time-trend coefficient with a *t*-statistic considerably less than 1.0. The frequent trips to the cafeteria windows (*S*'s received 60 or more payoffs per session), the low search costs involved in choosing within the confines of the experimental test chamber, and the physiological limits on substitutability imposed by the commodities employed all facilitate *S*'s adjusting rapidly to the changes in experimental conditions. The data here indicate that at least some of the subjects took advantage of these conditions.

The Durbin-Watson statistics, while indicating a tendency to negative first-order autocorrelation in the residuals, do not show a serious problem in these respects. Such a problem might be anticipated from the physiological interdependencies between food and water, which would result in the pigeon's getting caught short on food or water for a given day, only to be compensated for in the next day's consumption

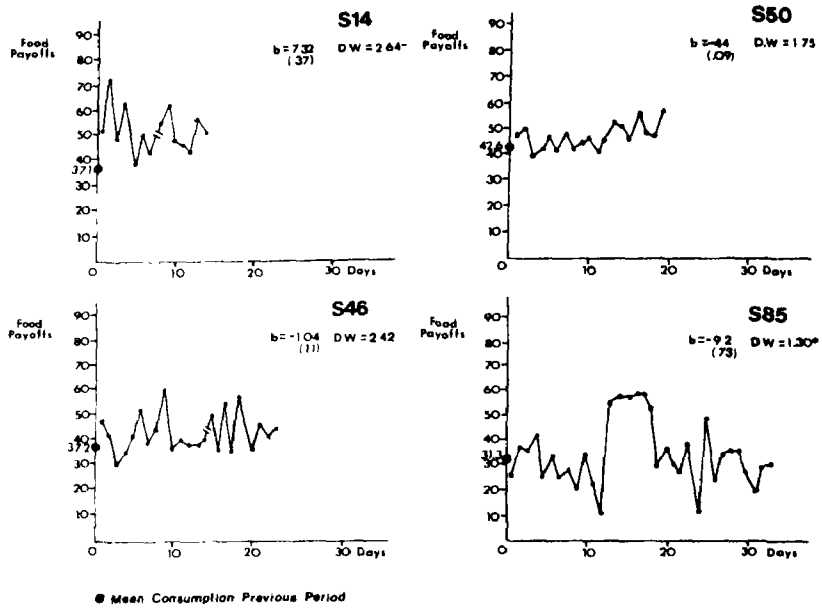


FIG. 2 — Temporal pattern of food consumption following a decrease in its relative price experiment 1, condition 4 Notation is the same as in figure 1.

bundle[4] The data indicate, however, that this did not happen. Subjects either had sufficiently good "planning," given the frequent trips to the cafeteria window in a given day, or sufficiently frequent crossovers between food and water schedules so as not to generate imbalances strong enough to dominate the residuals. Durbin-Watson statistics computed on the basis of fitting (8) to the data from the baseline periods show similar results.

III. Results for Experiment 2

Procedures

Essentially the same procedures were employed as in experiment 1. The major difference was the sequence of price changes.

Four experimentally naive male White Carneaux pigeons served as subjects. Initial (baseline) conditions were the same as for *S*'s 14, 46, and 50 in experiment 1 with the exception that *S* 38 obtained 13

[4] We are assuming that *S*'s can and will correct imbalances within one day. This assumption is based on data for days following equipment failures (e.g., a malfunctioning food dispenser) which show that, for rats and pigeons alike, excluding data for the day of the failure and the following day generally leaves the time trend unaffected.

seconds access with each food payoff and 6 seconds access to the water reservoir. The difference in price-change sequence from experiment 1 consisted of (1) decreasing relative food prices first and (2) not returning directly to baseline conditions but following this with a compensated increase in food prices relative to baseline before returning to baseline conditions. (Table 3 shows the sequence and magnitude of price changes.)

Comparative Static Results

Table 3 reports the effects of the compensated price changes. The decrease in relative food prices in period 2 led to statistically significant substitution effects in the expected direction in all cases. The price elasticities of demand for food ranged from a low of −.21 for *S* 38 to a high of −.51 for *S* 36, averaging −.32 across *S*'s. Applying *t*-tests to the percentage of total time spent obtaining food in periods 1 and 2, we again reject the hypothesis that the simple random-behavior model (7) explains any of the data: Demand is far too inelastic in this case also.

The compensated increase in relative food prices in period 3 does show some surprises, however. In all cases the relative price of food was greater than during the initial baseline period. However, the predicted magnitude of change in consumption did not materialize for any of the *S*'s. Although the composition of consumption (ratio of food to water payoffs) changed in the "right" direction from period 2 to 3 for three of the four *S*'s, in no case was the change sufficiently large to result in a correct sign for the weak-axiom test statistic (6). In contrast, comparable increases in relative food prices starting from baseline conditions in experiment 1 resulted in a correct sign for the test statistic in all cases. Further, a comparable increase in relative food prices for *S* 85 in period 2 of experiment 1, which started from a relative price ratio similar to that faced by *S*'s 36–39 in period 2 of this experiment, did result in consistent substitution effects significant at the 10 percent level.[5]

One potential explanation for the behavior observed in period 3 is that preferences had shifted. The return to baseline conditions in period 4 enables us to test this hypothesis. For the one subject, *S* 36, whose behavior is clearly inconsistent in period 3, preferences have clearly shifted ($t = 5.15$, df = 18, $P < .001$).[6] For the other three *S*'s showing nonexistent or undersubstitution effects, the hypothesis receives only weak support. Although food consumption has in-

[5] Real income was somewhat higher for *S* 85 in period 1.
[6] $P < .001$ = probability that we can reject the null hypothesis of no effect.

TABLE 3

RESULTS OF CONSISTENCY TESTS: EXPERIMENT 2

Subject and Experimental Condition	α_f/α_w	Weak-Axiom Test (t-Statistic)[a]	Mean Payoffs[b] (S_M) Food	Mean Payoffs[b] (S_M) Water	t-Test for Equality of Baselines (Marginal Significance Level)[c]
Subject 36					
1	1.0		24.1	35.9	
		−1.032	(1.26)	(1.26)	
		(11.3)***			
2	.22		47.5	29.7	
		.546	(1.89)	(.37)	5.15***
		(3.25)***			
3	9.02		27.6	23.3	(.001)
			(.37)	(3.15)	
4	1.0		33.2	26.8	
			(1.24)	(1.24)	
Subject 37					
1	1.0		41.2	18.8	
		−.312	(1.47)	(1.47)	
		(4.24)***			
2	.50		49.6	15.6	
		.318	(2.46)	(1.27)	.93
		(1.88)*			
3	4.0		42.1	12.6	(.37)
			(.90)	(3.71)	
4	1.0		43.8	16.2	
			(2.39)	(2.39)	
Subject 38					
1	1.0		45.0	15.1	
		−.228	(1.30)	(1.35)	
		(3.90)***			
2	.53		51.2	12.6	
		.54	(2.05)	(1.09)	.78
		(1.03)			
3	2.60		45.1	14.0	(.45)
			(.67)	(1.48)	
4	1.0		47.0	13.0	
			(2.21)	(2.21)	
Subject 39					
1	1.0		39.9	20.2	
		−.144	(2.00)	(2.01)	
		(2.51)**			
2	.53		47.0	15.4	
		0.0	(1.58)	(1.06)	1.78*
		(0.0)			
3	2.40		43.1	16.9	(.09)
			(.74)	(1.78)	
4	1.0		44.1	15.9	
			(1.25)	(1.25)	

NOTE.—α_f/α_w = relative food cost. S_M = standard error of the mean.

[a] The t-tests are two tailed with 9 degrees of freedom.

[b] Calculated over the last 10 days of each period.

[c] The t-tests are two tailed with 18 degrees of freedom.

* Significant at the 10 percent level.

** Significant at the 5 percent level.

*** Significant at the 1 percent level.

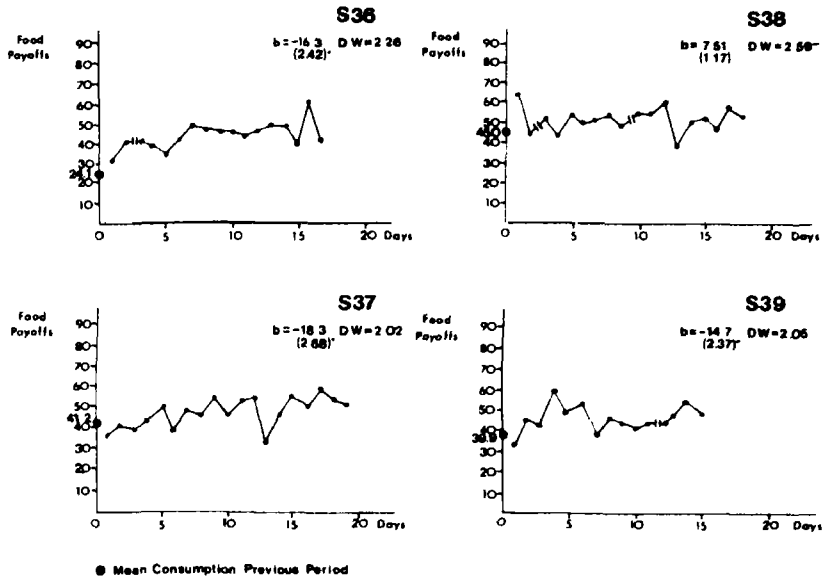


FIG. 3 — Temporal pattern of food consumption following a decrease in its relative price experiment 2, condition 2 Notation is the same as in figure 1

creased in period 4 relative to the initial period for all three, in no case does the increase seem to be sufficient, by itself, to explain the undersubstitution effects. The magnitude of the change in relative prices, and the fact that we are dealing with Slutsky-compensated price changes which, if anything, overestimate the Hicksian substitution effect (Friedman 1976), would seem to more than compensate for any small shifts in preference structure that might have occurred.[7] Finally, the behavior reported in period 3 is inconsistent with the random-response model (7) also.

Adjustment Patterns

Figure 3 shows food consumption in period 2 following the decrease in its relative price. The data show more of a positive time trend, consistent with the notion of habit formation, than comparable observations from experiment 1. Estimated values of the time-trend coefficient, b, from equation (8) confirm this as we have negative coefficients in three of the four cases, each one of which is significant at the 5 percent level. These negative coefficients indicate that subjects responded slowly to the decrease in relative food prices, consistent with notions of habit formation.

[7] We are assuming here that both food and water are normal goods.

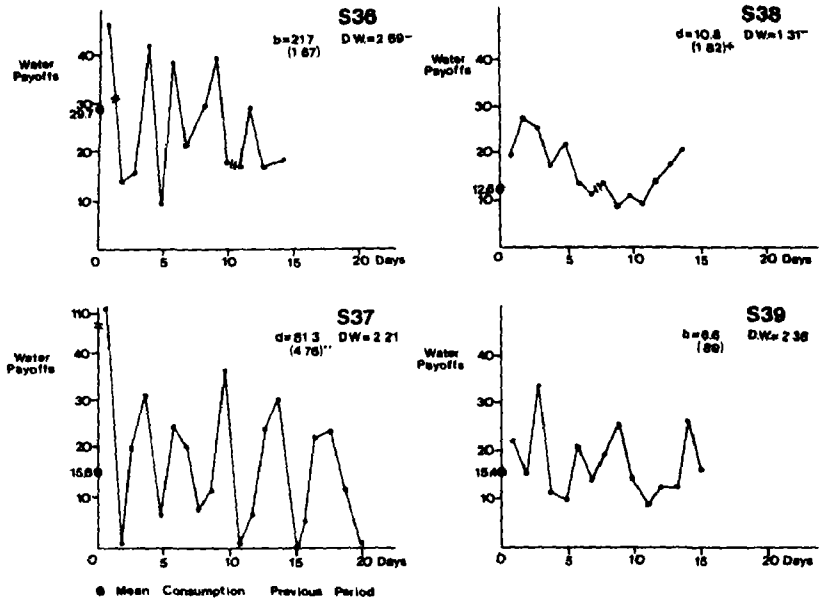


FIG. 4.—Temporal pattern of water consumption following a decrease in its relative price: experiment 2, condition 3. Notation is the same as in figure 1.

In contrast, the data for period 3, shown in figure 4, during which the relative price of water decreased, show stock-adjustment effects in all cases, being statistically significant at the 1 percent level for *S* 37 and at the 10 percent level for *S* 38. Thus, the initial response to the decrease in the price of water was greater at the beginning of this period than at the end. In fact, for all *S*'s the commodity bundle purchased on the first day of period 3 satisfied the basic consistency test (1). Further, for *S*'s 37–39, who did not shift preferences between baseline periods, the weak-axiom test statistic calculated over the first 3 days of period 3 has the correct sign, although it is somewhat smaller in magnitude than values observed under comparable conditions (period 2) in experiment 1. These data would appear to rule out arguments that the undersubstitution effects found in period 3 were a result of not maintaining experimental conditions for a sufficiently long period of time and/or that *S*'s failed to correctly process the magnitude of the changes in relative prices.

The adjustment patterns in the two experiments are qualitatively similar. Pooling the results from the two experiments shows an asymmetric response to increases and decreases in relative food prices. For decreases in prices we find a tendency to habit formation, or increased food consumption with the passage of time. For increases

in relative food prices the data consistently show anti–habit formation, more water (less food) consumed immediately following the price change than later on. These conclusions are reached whether we combine the t-test outcomes for the price-change periods ($Z = -2.28$, $P < .03$ for the t-values associated with b in periods when relative food prices decreased; $Z = 8.0, P < .001$ for the t-values associated with b in periods when relative water prices decreased) or use a χ^2 analysis based on the signs of the time-trend coefficients ($\chi^2 = 9.6$, df $= 1, P < .01$).[8]

The asymmetric response to increases and decreases in relative food prices is quite unexpected. One possible explanation lies in physiological asymmetries with respect to how animals process food and water. Food can be stored in the form of enhanced body weight, while the experimental procedures do not permit maintaining comparable stocks of water. Thus, in the case of increased food prices, there is a stock of food available to buffer large and immediate increases in water consumption. But with decreases in the price of food, there is not a stock of water to produce similar results. Further, food-buffer stocks were larger in experiment 2 at the time food prices increased since it came immediately after a large decrease in food prices in the previous period. This might have contributed to the stronger stock-adjustment effects found in this experiment.

Although adjustment patterns in the two experiments are qualitatively similar, a question arises as to what factors are responsible for the quantitative differences in the t-ratios and the size of the time-trend coefficients between the two experiments. The size of the reductions in relative food prices during condition 2 of experiment 2 is comparable to that faced by subjects in experiment 1 (see tables 1 and 3). Further, the length of the baseline periods preceding these price changes was similar in the two experiments. This would seem to rule out arguments that these quantitative differences result from the

[8] Under the hypothesis that the mean value for the t-statistic in the population is zero, the statistic,

$$z = \frac{\Sigma t_j}{\Sigma[f_j/(f_j - 2)]},$$

where f_j represents the degrees of freedom associated with t_j, has a sampling distribution which is approximately $N(0, 1)$ (Winer 1971). The picture gets a bit fuzzier if we include periods of return to baseline (conditions 3 and 5 in experiment 1, condition 4 in experiment 2) to the analysis. The Z-statistic for t-values when relative food prices decreased loses significance ($Z = -1.12, P = .26$), while for periods when relative water prices decreased the significance level drops ($Z = 2.60, P < .01$). Also, the significance level for the χ^2 test drops ($\chi^2 = 4.5$, df $= 1, P < .04$). The primary reason for this is that the sign of the b coefficient associated with food consumption in period 4 of experiment 1 is positive for S's 37–39, while to be more consistent with the price change periods it should be negative since food prices decreased going into this period.

magnitude of price changes or the length of time prices had remained stable. This leaves individual subject differences or differences in procedures (sequence effects or some undetected change) between the two experiments as the most likely factors responsible for the differences.

Attributing part of the differences to variability in individual subject responses receives some support. The learning-curve literature in psychology shows a lot of variability in response patterns using single-subject data, with some individual learning curves being very irregular or varying greatly from the general form for a given task (Mazur and Hastie 1978). Examination of the magnitude of the estimates of b within each experiment reveals a similar variability in behavior. For example, in both experiments one of the four subjects has a positive point estimate when the relative price of food decreased. Similar differences in individual subject responses are, of course, not detectable in aggregate per capita data such as Houthakker and Taylor employed. However, the consistent negative signs of the time-trend coefficients for decreases in relative food prices and positive signs for increases in relative food prices (decreases in the relative price of water) suggest that an aggregate per capita analysis, under conditions where all subjects faced the same relative price changes, would clearly exhibit these patterns.

Finally we note that the values for the Durbin-Watson statistics reported in figures 3 and 4 are comparable to those found in experiment 1. While there is a tendency to negative first-order autocorrelation in the residuals, these effects are rarely strong. Even though S's would readily consume more at the end of a session if given the opportunity, they managed to keep food and fluid intake relatively well balanced on a daily basis.

IV. Summary and Discussion

Using figure 5 we can readily summarize the consistency test results. Starting from some initial (baseline) budget line (line 1), with consumption bundle A, Slutsky-compensated increases or decreases in relative prices result in consistent and statistically significant substitution effects (points B and C, respectively). This result replicates similar tests of the Slutsky-Hicks theory using rats pressing levers and choosing between food and water, or flavored fluids with ad lib. access to food and water (Kagel et al. 1975, 1980). These results suggest the universality and robustness of one of the fundamental "truths" of the static theory of consumer demand, that compensated price increases will bring about a decrease in quantity demanded.

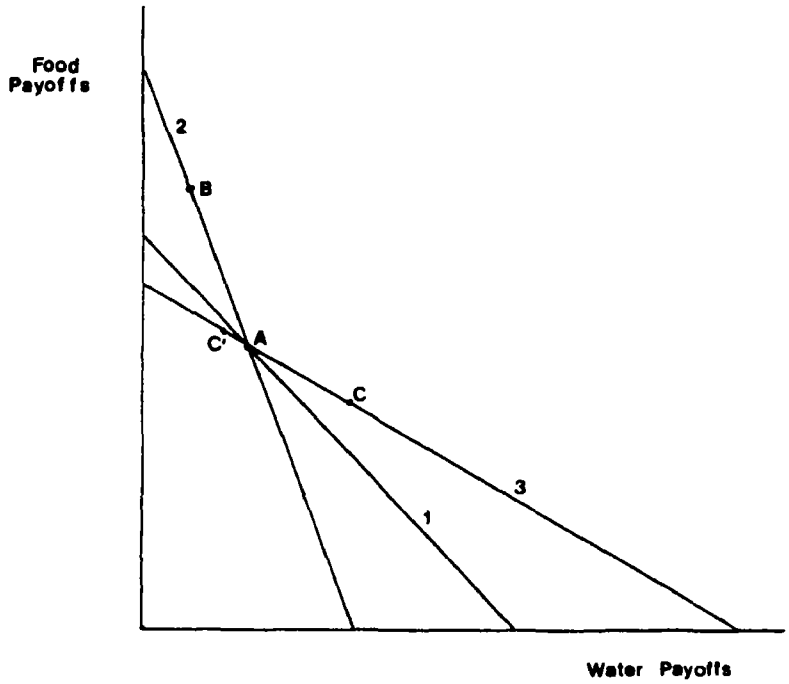


FIG. 5.—Summary of consistency-test results in terms of commodity-choice space. Numbered lines are budget lines under differing relative prices with capital letters characterizing choices. See text for details.

Further, this truth does not rest on simple random-behavior models suggested by Becker (1962) and Chant (1963) but apparently reflects a basic behavioral principle.

Large rotations in the budget line which involved swinging from line 2 to 3 without stopping off at baseline resulted in inconsistent choice behavior (point *C'*) relative to the initial commodity bundle consumed. While *S*'s tended to alter their composition of consumption in the expected direction, they did not change enough to be consistent with the initial commodity bundle consumed.[9] Efforts to explain these undersubstitution effects in terms of shifting preference structures and/or delayed reactions (learning) in response to the change in relative prices did not prove successful.

Although it would be nice to have an explanation for the inconsis-

[9] We searched our rat data for similar rotations in the budget line. In the one case this happened the results are similar to those reported here. This was overlooked at the time since further rotation of the budget line in the same direction resulted in additional changes in consumption so that the weak-axiom test statistic was satisfied.

tencies reported here in terms of established patterns of behavior derived from the choice experiment literature, we have not found one which is completely consistent. For example, anchoring effects, the fact that for humans estimates of quantities (Tversky and Kahneman 1974) and strengths of sensory phenomena (Stevens 1975) are greatly affected by the starting point used in making these estimates, might be used to explain the differences in going from budget line 2 to 3 directly rather than from 1 to 3. But anchoring, unless conjoined with a hypothesis specifying some minimum difference between points before it takes effect, would also seem to imply systematic biases in consumption patterns in returning to baseline conditions following a relative price change, something we did not observe in experiment 1. Explaining the inconsistencies in terms of changes in the animal's body weight resulting from the relative price changes runs into similar problems, although we cannot rule out more complex physiological interaction effects. Perhaps the behavior simply results from *S*'s inexperience, so that once having experienced a compensated price change starting from baseline prices, the same commodity bundle would be chosen starting from elsewhere in the choice space. Identification of the factors underlying the inconsistencies must wait further experimentation. Identifying these factors would also help clarify the behavioral mechanism underlying consistent choices.

Looking at day-to-day adjustments in consumption patterns resulting from changes in relative prices we find a number of cases in which adjustments were completed within the first observation period and other cases with statistically significant learning effects. Individual subject differences rather than the size of the price changes or other environmental factors were cited as one of the factors behind these differences. The fact that the adjustment coefficients tended to have the same sign whether they were statistically significant or not suggests that an aggregate analysis under conditions where all subjects faced the same price changes would produce statistically significant learning effects. Unanticipated asymmetries in the adjustment path resulting from increases and decreases in relative food prices were also found. We are looking at adjustment patterns from the rat studies in an effort to confirm these regularities.

Questions naturally arise concerning the relevance of the animal behavior reported here to understanding analogous human behavior, the behavior of primary interest to most economists. Objections might be raised on two grounds. One is that there are essential differences between the choice conditions (environmental conditions) the pigeons face and conditions facing human consumers operating in national economic systems. The pigeons are not given money (or

tokens, as in some human economic experiments) and told what prices are, but are simply faced with sequences of situations from which they must deduce what the rules of the game are. One might argue that this results in the pigeons facing a more difficult situation than is assumed in ordinary consumer-demand theory. Second, one might object that there are essential differences in the choice process itself, on what goes on inside the pigeon's brain, so to speak, relative to human consumers. The sequential way in which subjects learn about changes in relative prices and the apparent sequential nature of choices even under steady-state conditions stand in marked contrast to the textbook characterization of a consumer who maximizes a utility function subject to a budget constraint at a point in time, apparently deciding on all purchases at once. We deal with these objections below.

There is no question that the choice conditions the pigeons faced here have no exact physical analogue to choice conditions outside the laboratory for humans, or for pigeons in the wild for that matter. What an experiment aims for is to develop procedures in the laboratory which provide a logically consistent interpretation for the concepts and relationships of the theory under investigation. Within the context of consumer-demand theory, this involves developing procedures that alter the choice set in the same way that the symbols commonly designated "price" and "income" alter choices in the general theory. This is true of any application of the theory, be it experimental or nonexperimental, using human or nonhuman subjects, so that there are always some differences in environmental conditions in different applications of the theory, and these differences may affect decision outcomes.

If the theory is to have any generality it must hold under a variety of such choice conditions. Consequently, confirmations and disconfirmations are relevant to evaluating the truth status of the general theory under any and all logically consistent interpretations:[10] Further, it is often the case that the most preferred conditions for testing a theory will be "atypical." For example, one of the crucial experiments of relativity theory is to determine if stars behind the edge of the sun appear to be displaced from their known positions during a total solar eclipse. One does not object to such experimental observations simply on the grounds that they do not correspond to "normal" conditions (Smith 1980). Similarly, perfect compensation,

[10] Identifying differences in behavior under different interpretations of the general theory sets off a search for the factors responsible for these differences and eventually results in modifications of the general theory to account for these factors. Unfortunately, this is often an exceedingly slow process.

which is required to directly apply the weak-axiom test (1), rarely occurs in nature but has an important role to play in better understanding commodity-choice behavior. Establishing procedures for implementing these conditions using nonhumans is considerably less costly than with humans, while introducing tokens or some other conditioned reinforcer to serve as a medium of exchange and unit accounting, although technically feasible, would only complicate the experiment with little likelihood of seriously affecting behavior. In fact, we might argue that having subjects learn the rules of the game through experience rather than relying on price representations eliminates potential pyschological complications introduced by such representations in monetized economies (Russo, Krieser, and Miyashita 1975), proving a reference point against which to evaluate the predictions of the theory.

The apparent sequential nature of the pigeons' choices, while standing in contrast to notions of consumers possessing full information and making decisions simultaneously, poses no conflict with respect to observable characteristics of human choice in the marketplace. It takes real time for consumption to occur; hence, real choices of all organisms have a sequential aspect that is commonly ignored in comparative static analysis. Further, differences in "planning horizons" between humans and pigeons must, almost by definition, be one of length rather than existence, unless one is willing to argue that the pigeons' behavior is the result of some elaborate, genetically determined program.

Finally, it must be remembered that whether or not pigeons or humans have consciously thought out their behavior is irrelevant to characterizing that behavior as a solution to a constrained optimization problem. As Samuelson (1947) notes, "... it is possible to formulate our conditions of equilibrium as those of an extremum problem, even though it is admittedly not a case of an individual's behaving in a maximizing manner, just as it is often possible in classical dynamics to express the path of a particle as one which maximizes (minimizes) some quantity despite the fact that the particle is obviously not acting consciously or purposively." In a competitive world all organisms must solve the problem that time and energy are limited and that efficient allocation of resources, over which the individual has control, is essential to success. The fact that man is capable of self-awareness and reasoning and is occasionally guided by it in solving mathematical puzzles, for example, does not imply that these factors are either necessary or sufficient to achieving efficient allocation of resources. Nor does this prove that human consumers in the marketplace are guided by these factors, while pigeons in the laboratory are not.

Appendix

Procedural Details

The experiment was conducted in a sound-insulated pigeon test chamber that measured 13 ¼ inches long, 11 ½ inches wide, and 14 inches high. The test

TABLE A1

Subject and Experimental Condition*	No. of Days in Condition	α_f	α_w
Subject 14:			
1	5	60	60
2	12	72	25
3	46	60	60
4	16	34	120
5	14	60	60
Subject 36:			
1	20	60	60
2	19	20	90
3	16	120	13.3
4	56	60	60
Subject 37:			
1	46	60	60
2	19	45	90
3	20	80	20
4	27	60	60
Subject 38:			
1	18	60	60
2	20	48	90
3	15	72	27.7
4	48	60	60
Subject 39:			
1	49	60	60
2	17	48	90
3	15	72	30
4	31	60	60
Subject 46:			
1	14	60	60
2	12	80	28
3	39	60	60
4	25	26	120
5	28	60	60
Subject 50:			
1	8	60	60
2	11	80	20
3	50	60	60
4	19	38	120
5	19	60	60
Subject 85:			
1	16	36	72
2	22	90	20
3	33	60	60

NOTE.—α_f = average time between food payoffs in seconds, α_w = average time between water payoffs in seconds.
* Mean value reported in the text for S's 14 and 50 in condition 1 are over all days in this condition.

chamber had three response keys (2.5 centimeters in diameter, fabricated from clear acrylic plastic) located on the front panel where the food and water reservoirs were located. A single peck on the center response key switched modes of delivery from food to water and vice versa with no changeover delay. This key was transluminated with white light at all times. The other two response keys, located over the food hopper and water reservoirs, were color coded to indicate the delivery mode in operation. Responses on these keys did not affect the contingencies. General chamber illumination was provided by a 7-watt houselight. When commodities were delivered, the response keys were darkened but the houselight remained on. A fan provided ventilation, and constant white masking noise was provided.

Food and water deliveries were programmed by separate variable time tapes. The tapes established a random sequence of interreinforcement times following an exponential distribution under all conditions. The sequence of interreinforcement times repeated itself every 20–30 reinforcements.

Food reinforcements occurred only when in the food mode and consisted of the food hopper coming into reach and the food hopper lights coming on during reinforcement. Water reinforcement occurred only when in the water mode and consisted of pumping water into a reservoir at the rate of 1 milliliter per 15 seconds and operating a response feedback clicker every ½ second. During reinforcement, pecks on the center response key did not result in a mode switch.

Table A1 shows the number of days each condition was in effect and the average absolute time delays (in seconds) between food and water payoffs.

References

Becker, Gary S. "Irrational Behavior and Economic Theory." *J.P.E.* 70, no. 1 (February 1962): 1–13.

———. *Economic Theory*. New York: Knapp, 1971.

Chant, John F. "Irrational Behavior and Economic Theory: A Comment." *J.P.E.* 71, no. 5 (October 1963): 505–10.

Friedman, Milton. *Price Theory*. Chicago: Aldine, 1976.

Hicks, John R. *Value and Capital*. 2d ed. Oxford: Clarendon, 1946.

———. *A Revision of Demand Theory*. London: Oxford Univ. Press, 1956.

Hirshleifer, Jack. "Natural Economy versus Political Economy." *J. Soc. and Biological Structures* 1 (October 1978): 319–37.

Houthakker, Hendrik S., and Taylor, Lester D. *Consumer Demand in the United States: Analyses and Projections*. 2d ed. Cambridge, Mass.: Harvard Univ. Press, 1970.

Kagel, John H., and Battalio, Raymond C. "Token Economy and Animal Models for the Experimental Analysis of Economic Behavior." In *Evaluation of Econometric Models*, edited by Jan Kmenta and James B. Ramsey. New York: Academic Press, 1980.

Kagel, John H.; Battalio, Raymond C.; Green, Leonard; and Rachlin, Howard. "Consumer Demand Theory Applied to Choice Behavior of Rats." In *Limits to Action: The Allocation of Individual Behavior*, edited by John E. R. Staddon. New York: Academic Press, 1980.

Kagel, John H.; Battalio, Raymond C.; Rachlin, Howard; Green, Leonard; Basmann, Robert L.; and Klemm, W. R. "Experimental Studies of Consumer Demand Behavior Using Laboratory Animals." *Econ. Inquiry* 13 (March 1975): 22–38.

Mazur, James E., and Hastie, Reid. "Learning as Accumulation: A Reexamination of the Learning Curve." *Psychological Bull.* 85 (November 1978): 1256–74.

Phlips, Louis. *Applied Consumption Analysis.* New York: American Elsevier, 1974.

Rachlin, Howard. "Economics and Behavioral Psychology." In *Limits to Action: The Allocation of Individual Behavior*, edited by John E. R. Staddon. New York: Academic Press, 1980.

Rapport, David J., and Turner, James E. "Economic Models in Ecology." *Science* 195 (January 28, 1977): 367–73.

Rees, Albert. "Economics." In *International Encyclopedia of the Social Sciences*, edited by David L. Sills. New York: Macmillan and Free Press, 1968.

Russo, J. Edward; Krieser, G.; and Miyashita, S. "An Effective Display of Unit Price Information." *J. Marketing* 39 (April 1975): 11–19.

Samuelson, Paul A. *Foundations of Economic Analysis.* Cambridge, Mass.: Harvard Univ. Press, 1947.

Smith, Vernon L. "Relevance of Laboratory Experiments to Testing Resource Allocation Theory." In *Evaluation of Econometric Models*, edited by Jan Kmenta and James B. Ramsey. New York: Academic Press, 1980.

Stevens, Stanley S. *Psychophysics: Introduction to Its Perceptual, Neural, and Social Prospects*, edited by Geraldine Stevens. New York: Wiley, 1975.

Tversky, Amos, and Kahneman, Daniel. "Judgement under Uncertainty: Heuristics and Biases." *Science* 185 (September 27, 1974): 1124–31.

Winer, Ben J. *Statistical Principles in Experimental Design.* 2d ed. New York: McGraw-Hill, 1971.

Taste Change in the United Kingdom, 1900–1955

Steven E. Landsburg

University of Chicago

A search is made for violations of the axiom of revealed preference in aggregate consumption data for the United Kingdom. No such violations are found. This is taken as evidence that tastes remained constant throughout the period under study. The strength of this evidence is estimated within the context of a particular model of taste change.

I. Introduction

This paper investigates the phenomenon of "taste change" in the United Kingdom in the years 1900–1955. In the paragraphs which follow we shall describe a method of searching for evidence of that phenomenon. In Section II we shall carry out several versions of that search and find no such evidence. In Section III we shall develop a general model of taste change in the context of which we shall strengthen the negative result of Section II. Section IV is a further discussion of the model. In Section V we shall discuss some of the implications of our results for economic analysis.

Our basic methodology is similar to that of Houthakker (1963) and Koo (1963). First we choose a set of goods. Then we choose a pair of years—say, for concreteness, 1900 and 1910. We observe what quantity of each good was consumed in 1900 (we shall refer to this set of quantities as the 1900 market basket), and we calculate what the price of this market basket would have been had it been purchased in 1910. If the price is lower than the price of the basket which was actually consumed in 1910, then we conclude that (in 1910) the 1910 basket was considered preferable to the 1900 basket. We repeat this procedure for every possible pair of years in some time period. Then we

[*Journal of Political Economy*, 1981, vol. 89, no. 1]
© 1981 by The University of Chicago. 0022-3808/81/8901-0006$01.50

search for intransitivities—for example, the 1910 basket preferred to the 1900 basket, the 1900 basket preferred to the 1905 basket, and the 1905 basket preferred to the 1910 basket. In other words, we search for violations of the Strong Axiom of Revealed Preference. Such an intransitivity is taken as evidence of a change in tastes.

These methods can be applied either to total national consumption or to per capita consumption. We have done both.

A problem arises when a good is available only in certain years—for example, there appear to have been no raspberries in Britain before 1922. This problem (which was rare) was solved by declaring the good in question to be identical with some other good, the choice of "other good" being dictated partly by aesthetic considerations and partly by the observation that identical goods should have nearly identical price series. (Raspberries were classed with black currants.)

When a good is included in the market basket, it is important to include all close complements and substitutes for that good as well. If, for example, our basket consisted only of automobiles and bicycles, we might discover a preference for more bicycles and fewer automobiles in 1978 than in 1971—reflecting not a taste change but a rise in the price of gasoline. The problem would not arise if gasoline were included in the basket.

II. Evidence

The methods described in Section I were applied to six sets of data, which are described below. No intransitivities were observed in any case.

We first considered total national consumption of nine goods: food, alcoholic beverages, tobacco, clothing, transport and communication, housing, fuel, household durables, and "other." The study covers the years 1900–1955. It is noteworthy that no intransitivities occur throughout this period despite the fact that the definition of the United Kingdom changed with the secession of Ireland in 1920. This may be taken as a small bit of evidence that the tastes of the nation did not change even when the composition of the nation did.

The second study was identical to the first with respect to the time period and the goods considered; it reflected per capita rather than total consumption.

The third and fourth studies were identical to the first and second with the exception that housing was eliminated from consideration. This was because expenditure on housing was defined as the total of rents, rates, and water charges collected, a definition which is clearly objectionable (e.g., it fails to include the opportunity cost of owner-occupied housing).

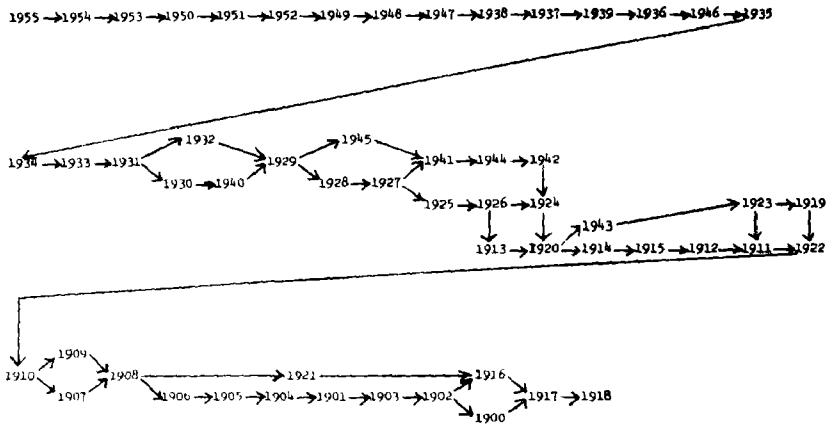


FIG. 1.—Revealed preferences among years for total U.K. consumption of all items, classified into nine categories.

Finally, we broke food, alcohol, and tobacco down into 107 subcategories and studied data on those 107 goods for the years 1920–38. Again this was done both for total and per capita expenditures.

Figure 1 provides an example of the results which were obtained. It depicts the conclusions of the first study described above, concerning total national consumption of nine broadly defined categories of goods. An arrow $A \rightarrow B$ indicates that the consumption of year A was revealed preferred to that of year B. Revealed preferences which can be inferred by transitivity from the ones depicted have been omitted from the figure in the interests of clarity. Analogous charts for the other studies described above are available from the author.

The data for the first four studies mentioned were taken from Mitchell and Deane (1954); for the other two they were taken from Stone and Rowe (1954).

It should be noted that the upward trend in income, particularly toward the end of the period under consideration, would tend to hide changes in taste even if they did occur.[1] However, the tests in the next section will explicitly take this phenomenon into account.

III. Interpretation

Our failure to observe a taste change supports the belief that no taste change occurred; now we ask how strongly that belief can be supported. Upon making certain assumptions and reexamining the data, we shall be able to infer that it is unlikely that the belief is far wrong.

[1] I am indebted to Gary Becker for this observation.

While the model which we shall develop could theoretically be applied to a market with an arbitrary number of goods, practical considerations such as limited computational resources require that we apply it only in the two-good case. For that reason, we shall restrict our detailed description of the model to the two-good case; the results for the n-good case are available on request.

Let X and Y be two goods with prices p_X and p_Y. We shall suppose that the utility function is the "constant elasticity of substitution" function $U(X,Y) = C(AX^u + BY^u)^{1/u}$, where A, B, C, and u are constants and $0 \neq u < 1$. It follows that the elasticity of substitution between X and Y is everywhere equal to $1/(1 - u)$.

We have $dU/dX = AC(AX^u + BY^u)^{(1/u)-1}X^{u-1}$ and $dU/dY = BC(AX^u + BY^u)^{(1/u)-1}Y^{u-1}$. By changing the units in which X and Y are measured we may assume that utility is maximized when one unit of each is consumed, from which we infer that $(dU/dX)/(dU/dY)\big|_{(1,1)} = p_X/p_Y$, that is, $A/B = p_X/p_Y$.

We suppose for the moment that tastes change while relative prices and real income do not. We shall interpret a change in tastes as a shift to a new utility function $V(X,Y)$ which is such that $dV/dX = \alpha(dU/dX)$ and $dV/dY = \beta(dV/dY)$, where α and β are observations of random variables $\hat{\alpha}$ and $\hat{\beta}$. That is, we assume that at the margin, any unit of X or Y will yield α or β times as much utility as it did before. The possibilities $\alpha \leqslant \beta$ and $\beta \leqslant \alpha$ should be equally likely, so we assume that $\hat{\alpha}/\hat{\beta}$ has a median of one. The function satisfying these requirements is $V(X,Y) = C(\alpha AX^u + \beta BY^u)^{1/u}$. This has the same indifference curves as $V(X,Y) = C[(\alpha/\beta)AX^u + BY^u]^{1/u}$, so we may replace α by α/β and β by one. Thus our assumption is that the new utility function has the form $V(X,Y) = C(\alpha AX^u + BY^u)^{1/u}$, where α is an observation of a random variable $\hat{\alpha}$ with a median of one. (The hypothesis that tastes do not change is the hypothesis that $\hat{\alpha}$ has variance zero.)

Now $dV/dX = \alpha AC(A\alpha X^u + BY^u)^{(1/u)-1}X^{u-1}$ and $dV/dY = BC(A\alpha X^u + BY^u)^{(1/u)-1}Y^{u-1}$. In light of the budget constraint, the new levels of consumption must be $1 + t$ units of X and $1 - (A/B)t$ units of Y for some number t. In order to maximize utility, t must satisfy

$$\frac{dV/dX}{dV/dY}\bigg|_{(1+t,\,1-(A/B)t)} = p_X/p_Y = A/B,$$

that is,

$$\frac{\alpha A(1 + t)^{u-1}}{B[1 - (A/B)t]^{u-1}} = A/B$$

or

$$t = \frac{1 - \alpha^{1/(u-1)}}{(A/B) + \alpha^{1/(u-1)}}.$$

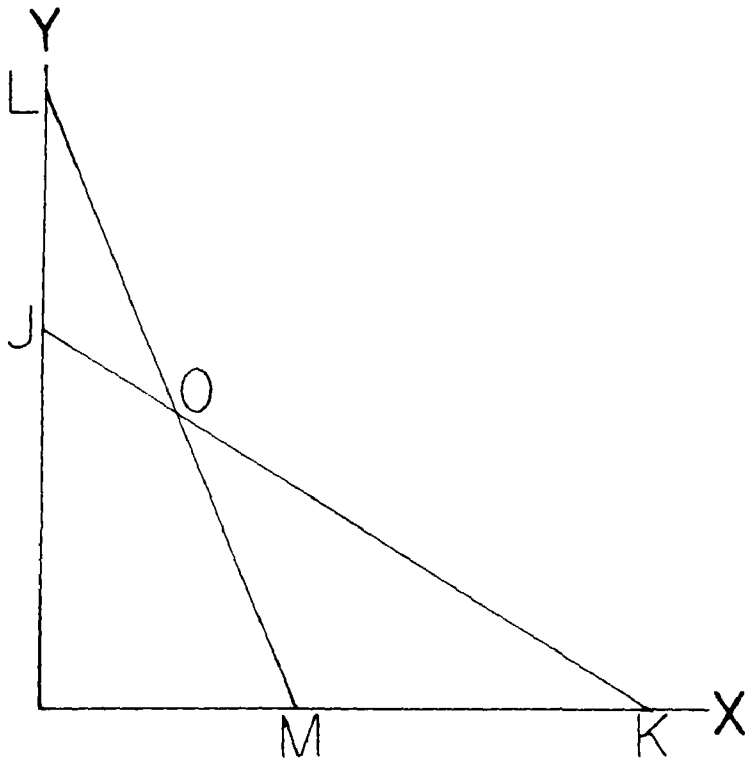


FIG. 2

Now let us suppose that in our second period we have not only a new utility function but a new budget line as well. If the two budget lines fail to cross, then there is no hope of observing a taste change, because no point on the inner budget line can ever be revealed preferred to any point on the outer one. Thus we restrict our attention to the case in which the budget lines *do* cross, as in figure 2. We can rename the goods so that the price of X falls relative to that of Y between the first and second periods; thus we may assume that LM represents the first-period budget line and JK the second-period budget line.

When the basket consumed in the first period is on segment OM and the basket consumed in the second period is on segment OJ, each is revealed preferred to the other and a taste change is observed. An estimate of the probability of this occurrence would yield a measure of the strength of our negative result; however, we shall argue that the hypothesis of taste change can be more strongly rejected through consideration of the "exact opposite" situation—that in which neither

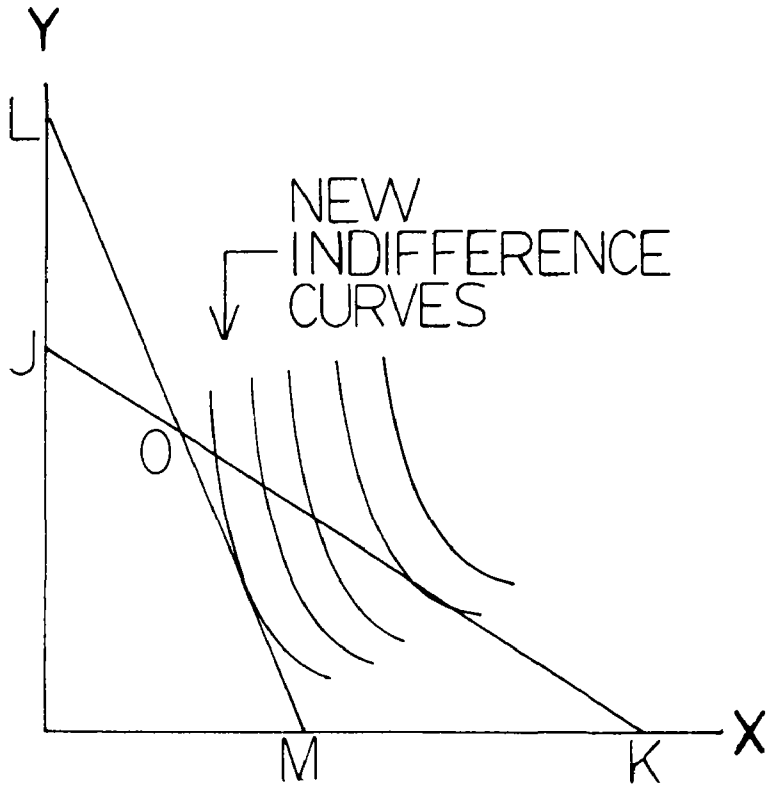


FIG. 3.—If there is an indifference curve tangent to OM, there must be one tangent to OK

basket is revealed preferred to the other. In terms of figure 2, this situation (which we shall refer to as incomparability of the baskets) arises when the first period's basket is on LO and the second period's basket is on OK.

Now suppose that the first period's basket is observed to be on LO. In keeping with our earlier assumptions, our supposition is that the point (1,1) lies on LO. If the utility function had changed and the budget line had not, the new utility-maximizing market basket would have been at the point $[1 + t, 1 - (A/B)t]$ on line LM. If this point happens to be on OM, then with the new budget line the utility-maximizing market basket will be on OK, as is evident from figure 3.

It follows that if we know that the first period's market basket lies on LO and the point $[1 + t, 1 - (A/B)t]$ lies on OM, we will observe an incomparability.

Denoting the coordinates of 0 by X_0 and Y_0, we see that $[1 + t, 1 - (A/B)t]$ is on OM if and only if $1 + t \geqslant X_0$. Since

$$t = \frac{1 - \alpha^{1/(u-1)}}{(A/B) + \alpha^{1/(u-1)}},$$

this is the case if and only if

$$\frac{1 - \alpha^{1/(u-1)}}{(A/B) + \alpha^{1/(u-1)}} \geqslant X_0 - 1,$$

that is,

$$\frac{1 + (1 - X_0)(A/B)}{X_0} \geqslant \alpha^{1/(u-1)}$$

or—since $A/B = (Y_0 - 1)/(1 - X_0)$—$Y_0/X_0 \geqslant \alpha^{1/(u-1)} = \alpha^{-\eta}$, where η is the elasticity of substitution between X and Y.

Conclusion

If, in figure 3, the first period's market basket is observed to be on LO (i.e., outside the budget constraint of the second period), and if Y_0/X_0 is greater than $\alpha^{-\eta}$, then we will observe an incomparability, provided X and Y are measured in units such that the first period's market basket is at the point (1,1).

Similar considerations show that if the second period's market basket is observed to be on OK and if units are chosen so that its coordinates are (1,1), then an incomparability will be observed provided that $X_0/Y_0 \geqslant \alpha^{-\eta}$.

Remarks

1. Because we shall be applying our model only in the two-good case, we shall have to assume that consumption of all substitutes and complements for the goods in question remains relatively constant (this is so in light of the discussion at the end of Section I). We shall approximate this condition by requiring that the first and second periods be successive years.

2. While the form of the utility function precludes the possibility that $\eta = 1$, we shall assume for ease of exposition that it is very close to one. (In Section IV we will explore the effect of letting η take on other values.) Thus we reinterpret the conclusion as saying that an incomparability will be observed provided Y_0/X_0 or X_0/Y_0, whichever is appropriate, is greater than α^{-1}.

The data which were used to construct figure 1 were reexamined, and values of Y_0/X_0 or X_0/Y_0 were computed for every pair of

successive years and every pair of goods with crossing budget lines. The results, for total and per capita consumption, with and without the questionable housing data, are displayed in table 1 at the end of this section. (In these new studies, as in the earlier ones which involved more than two goods at a time, no intransitivities were observed.)

Under total consumption there are 604 relevant cases (i.e., cases where the budget lines cross). Only nine incomparabilities were observed. More important, there are two cases in which an incomparability would have been observed if α^{-1} had been less than 1.00. If we assume that tastes *do* change (i.e., that $\hat{\alpha}$ is not identically one), then, by our assumption that $\hat{\alpha}$ has median one, α^{-1} must be less than one 50 percent of the time; therefore the probability of our observing an incomparability in one of these two cases is greater than $1 - (1 - 0.5)^2 = 0.75$. Thus our failure to observe an incomparability in either case allows us to reject the hypothesis of taste change with 75 percent confidence. The above remarks apply to total consumption with or without housing data.

Under per capita consumption, there are 673 cases and only three incomparabilities. There are 14 cases in which an incomparability would have been observed if α^{-1} had been less than 0.95. (There are eight such cases if one ignores the housing data.) Among these cases there is one incomparability. If we were to assume that α^{-1} is less than 0.95 even 25 percent of the time, the probability of its being so only once in 14 trials would be only 10.1 percent. Thus we can say with 89.9 percent confidence (63.3 percent confidence if we do not use the housing data) that α^{-1} is not less than 0.95 even as much as 25 percent of the time.

In three cases, none of which involved housing data, X_0/Y_0 or Y_0/X_0 was greater than 0.995. No incomparabilities were observed in any of these cases. Calculating as above, we find that this allows us to reject, with 57.8 percent confidence, the hypothesis that α^{-1} is less than 0.995 more than 25 percent of the time.

Summary of Results

Using total consumption, α^{-1} never differs from one (75 percent confidence). Using per capita consumption, α^{-1} is not less than 0.95 more than 25 percent of the time (89.9 percent confidence, 63.3 percent without housing), and α^{-1} is not less than 0.995 more than 25 percent of the time (57.8 percent confidence). These results say, with some degree of confidence, that $\hat{\alpha}$ seems to have very small variance—and this is just what is meant by the assertion that tastes do not change.

TABLE 1

X_0/Y_0 or Y_0/X_0	With Housing Data		Without Housing Data	
	No. of Cases	No. of Incomparabilities	No. of Cases	No. of Incomparabilities
Total consumption.				
.00-.049	54	0	46	0
.05-.099	40	0	31	0
.10-.149	43	0	32	0
.15-.199	39	0	27	0
.20-.249	54	0	43	0
.25-.299	45	0	33	0
.30-.349	34	0	26	0
.35-.399	30	0	21	0
.40-.449	22	0	16	0
.45-.499	28	0	22	0
.50-.549	33	0	27	0
.55-.599	20	0	15	0
.60-.649	28	0	21	0
.65-.699	23	0	14	0
.70-.749	25	0	20	0
.75-.799	21	0	17	0
.80-.849	20	1	13	1
.85-.899	13	0	11	0
.90-.949	15	0	8	0
.95-.999	15	8	10	7
1.00	2	0	2	0
Total	604	9	455	8
Per capita consumption				
.00-.049	62	0	47	0
.05-.099	44	0	38	0
.10-.149	52	0	40	0
.15-.199	56	0	37	0
.20-.249	51	0	39	0
.25-.299	42	0	30	0
.30-.349	39	0	22	0
.35-.399	32	0	25	0
.40-.449	34	0	25	0
.45-.499	32	0	23	0
.50-.549	30	0	25	0
.55-.599	26	0	21	0
.60-.649	29	0	19	0
.65-.699	29	0	21	0
.70-.749	22	0	17	0
.75-.799	21	0	11	0
.80-.849	24	0	20	0
.85-.899	15	0	9	0
.90-.949	19	2	13	1
.95-.999	13	1	7	1
1.00	1	0	1	0
Total	673	3	500	2

It is important to note that the results listed above were derived without any reference to the fact that in all of our studies we never observed a taste change (the phrase "all of our studies" refers both to the set of studies described in Section II and to the 1,277 separate studies described in this section). This striking fact, which is really the main result of this paper, surely mitigates our disappointingly low levels of confidence, but I know of no way in which to quantify the extent to which it does so.[2]

IV. Discussion

Like most models, ours is laden with assumptions with which a reasonable man might take issue. Here we reconsider some of the points which are most open to question.

1. *The Form of the Utility Function*

It is certainly the case that our choice is arbitrary. However, utility functions of this form are often assumed in economics, and our results do indicate that we should be wary of models which assume such utility functions and also allow for changes in taste. Thus, even if we were to grant that our utility function is not an accurate reflection of economic reality, we would still be able to draw conclusions about the consistency of certain economic models, and this is a topic which should be of concern to economists.

It should also be noted that, after some effort, I have not been able to find a "reasonable" utility function which yields results substantially different than does the one adopted here.

[2] One way to begin to quantify it is the following. Let p_X^i and p_Y^i be the prices of X and Y in the ith year and let

$$D = \frac{p_X^2 p_Y^1}{p_X^1 p_Y^2}$$

(so that D is always less than one, because the goods were renamed so that the relative price of X fell). Then it can be shown that a taste change will be observed if and only if $\alpha < (X_0/Y_0)D$ or $\alpha < (Y_0/X_0)D$ (depending on whether the first year's basket is inside the second year's budget line or vice versa). Now we could proceed as in the body of the paper, measuring values for $(X_0/Y_0)D$ and $(Y_0/X_0)D$ instead of X_0/Y_0 and Y_0/X_0. This procedure—which amounts to measuring the strength of our failure to observe a taste change by measuring the probability that a taste change could occur and go unobserved—seems more straightforward than the one adopted in the text, but when applied to the same data it yields numerical results which are substantially weaker. Given the assumptions of our model, this means that even if tastes changed substantially, our failure to observe a taste change in any of the 1,277 cases of Section III would not be too surprising. However, this calculation ignores all of the findings of Section II, which appear to be quite strong. I do not know the best way to quantify their strength—a naive generalization of the current model would require additional assumptions and very extensive computational resources.

2. *The Exogeneity of Price Changes*

On the surface this is a very damaging objection. If prices and tastes do not change independently, as we have assumed, than a considerably subtler analysis of the problem is required. However, reasoning as in (1) above, we argue that the fact that many economic models do assume exogenous price changes implies that we can at least draw a conclusion about those models. The relevant question for economists is not "Do tastes change?" but "Are models which assume taste changes consistent with reality?" The latter is the question which we have really been addressing; the answer, for a large class of models, appears to be "no."

3. *The Form of the Taste Change*

All that can really be said here is that the form which we have assumed seems reasonable and that other reasonable forms yield similar results. If, for example, we had assumed that the new utility function was given by $V(X,Y) = U(\alpha X,Y)$, rather than being determined by the requirement that $dV/dX = \alpha(dU/dX)$, then our conclusion would have been exactly the same except that $\alpha^{1-\eta}$ would have appeared instead of $\alpha^{-\eta}$.

4. *The Assumption That η Is Close to One*

Our calculation of the confidence with which we can say that α never differs from one (i.e., that tastes do not change at all) is unaffected by this assumption: It would remain the same for any value of η. With regard to our calculation of the confidence with which we can say that α is not less than some given value more than a given percentage of the time, our result would be stronger for larger η and weaker for smaller η.

5. *The Assumption That α Is Drawn Independently Each Time*

This assumption was essential when we calculated confidence levels in Section III. Our first calculation was that we could say with 75 percent confidence that α never differs from one. The two cases involved in this calculation were transport and communication versus tobacco for 1907–8 and fuel versus alcoholic beverages for 1937–38. The 30-year separation, together with the fact that entirely different goods are involved, seems to justify the assumption that α was drawn independently in these two cases.

Our strongest numerical result involved the confidence with which we could say that α was not less than 0.95 more than 25 percent of the

time; it was calculated to be 89.9 percent on the basis of 14 cases. The 14 cases involved are listed in the unnumbered table below.

1. Housing vs. food	1913–14
2. Housing vs. alcoholic beverages	1942–43
3. Housing vs. tobacco	1916–17
4. Housing vs. tobacco	1943–44
5. Housing vs. fuel	1951–52
6. Household durables vs. tobacco	1922–23
7. Household durables vs fuel	1927–28
8. Household durables vs. fuel	1929–30
9. Household durables vs. transport and communication	1905–6
10. Household durables vs. transport and communication	1909–10
11. Alcoholic beverages vs. other	1909–10
12. Household durables vs. other	1904–5
13 Housing vs. other	1902–3
14. Fuel vs. other	1901–2

(Case 14 was the one in which an incomparability was observed.) The reader is invited to form his own opinion as to how many of these cases are really independent. If, for example, he discards cases 8 and 10 on the grounds that they are too closely tied to cases 7 and 9, then we have 12 independent cases instead of 14 and our confidence drops from 89.9 percent to 84.2 percent.

V. Conclusions

The main result of this paper is that no taste change was observed in any of our studies. These included studies of total consumption and of per capita consumption, studies of broadly defined categories of goods and of very narrowly defined goods, studies of exhaustive and nonexhaustive lists of goods, and studies of as many as 107 and as few as two different goods. The two-good studies alone presented 1,980 separate cases; 1,277 of these had crossing budget lines and hence the potential for an observed taste change. None of this potential was realized. In Section III we calculated levels of confidence with which we could say that tastes did not change by more than certain specified amounts; these calculations were independent of the fact that no taste change was observed and therefore greatly understate the strength of our results.

The morals are the obvious ones: In economics, we should be wary of models which include variables for taste change, and outside economics we should be skeptical of warnings about the "wild vicissitudes of taste" and the constant danger of having our vicissitudes manipulated. With regard to this last point, it is particularly interesting to note that our study covers the period in which the electronic media came into being and into flower.

It is easy to postulate specific forms of taste change among the goods which we have studied, and it is easy to postulate forms which our methods could not have observed. It is considerably harder to postulate such forms which satisfy the additional restriction that they be reasonable. I do not believe that is possible, at least not without violating exogeneity of prices. If this is correct, then our results have strong implications for all price-theoretic models in which price changes are exogenous.

References

Houthakker, Hendrik S. "Some Problems in the International Comparison of Consumption Patterns." In *Les Besoins de biens de consommation*, an international colloquium of the Centre National de la Recherche Scientifique. Grenoble: Centre National de la Recherche Scientifique, 1963.

Koo, Anthony Y. C. "An Empirical Test of Revealed Preference Theory." *Econometrica* 31 (October 1963): 646–64.

Mitchell, Brian R., and Deane, Phyllis. *Abstract of British Historical Statistics*. Cambridge: Cambridge Univ. Press, 1954.

Stone, J. Richard, and Rowe, D. A. *The Measurement of Consumers' Expenditure and Behaviour in the United Kingdom, 1920–38*. Cambridge: Cambridge Univ. Press, 1954.

The Determinants of Tariff and Nontariff Trade Restrictions in the United States

Edward John Ray

Ohio State University

This paper develops and tests a simple model for the determination of tariff and nontariff barriers to trade across industries within the United States, using 1970 trade data. We find that nontariff trade restrictions have supplemented tariff protection in the United States. Both tariff and nontariff trade restrictions are biased toward industries in which the United States has an apparent comparative disadvantage in world trade and away from industries in which consumer welfare losses from protection would be great. We also find substantial evidence that tariff and nontariff trade restrictions predominate in industries with very different market characteristics.

The purpose of this study is to synthesize a number of ideas expressed in earlier work into a simple analytical framework to explain the structure of tariff and nontariff barriers to trade across industries in the United States. Our central premise is that, subject to political constraints, trade restrictions are consistent with profit maximization across industries. Using International Trade Commission data for tariff and nontariff trade restrictions and industry characteristics for 225 U.S. manufacturing industries in 1970, we find that such an approach has significant explanatory power.

The resulting empirical analysis is unique in a number of important respects. First, we provide separate reduced form estimates of tariff

The author is grateful to Betsy Radigan, Steve Cook, Ed Honton, and Jim Brown for aid in generating the Tobit and Probit results. Professor Tetsunori Koizumi, Jacob Frenkel, Howard Marvel, Donald Parsons, Jerry Thursby, Marie Thursby, Thomas Wolf, and, most particularly, J David Richardson provided numerous helpful comments on earlier drafts of this paper.

Journal of Political Economy, 1981, vol 89, no 1]
© 1981 by The University of Chicago 0022-3808/81/8901-0009$01 50

and nontariff trade restrictions in the United States. Second, by extending our analysis to include the estimation of a simultaneous model for the determination of tariff and nontariff trade restrictions, we provide empirical support for our hypothesis that nontariff trade restrictions have been utilized in part to compensate for internationally agreed-upon tariff reductions in the post–World War II period.

The third important finding is that, while tariff and nontariff trade restrictions co-exist in industries with some common characteristics, there are some distinct differences in industry characteristics that are more strongly associated with one rather than the other alternative form of protection. Specifically, we find that both tariff and nontariff trade restrictions are biased toward industries in which the United States is at a comparative disadvantage in world trade, and away from industries in which losses in consumer welfare from protection would be particularly large. But there are some real differences too. Tariffs are biased toward industries which are low-skill intensive, away from industries that are capital intensive, and unrelated to product heterogeneity, the concentration of production, and the geographical dispersion of domestic production facilities. In contrast, nontariff trade restrictions are concentrated in industries producing fairly homogeneous products using relatively capital-intensive techniques of production that, at the same time, are not intensive users of low-skilled labor. In addition, nontariff trade restrictions are found predominantly in industries in which production is less concentrated and in which production facilities are distributed across regions of the United States in a fashion consistent with the distribution of population and, therefore, voting power in Congress.

I. The Political Economy of Trade Restrictions

The general approach of this paper is that the structure of tariff and/or nontariff trade restrictions across industries is consistent with the simple joint maximization of industry profits subject to political constraints. In effect, we assume that individual industry characteristics influence the profitability of trade restrictions and that political factors similar to those discussed by Cheh (1974, 1976), Pincus (1975), Baldwin (1976*a*), Caves (1976), Caves and Jones (1977), Helleiner (1977), Stone (1978), and others are critical in determining which industries are most successful in obtaining restrictions on imports of competitive products.

The primary information to emerge from a simple analysis of the relationship between industry rents and protection is that the profitability of trade restrictions should normally be negatively related to the absolute value of the own price elasticity of demand for the product

and positively related to the foreign elasticity of supply of the product.[1] In effect, the more responsive consumers are to price increases induced by trade restrictions and the less responsive foreign suppliers are to changes in the international price of the good, the less profitable any given level of protection will be.

Furthermore, a discussion of the profitability of protection assumes that imports represent a reasonable threat to sales in the United States by domestic producers. If concern for industry profits is important in explaining the structure of trade restrictions, we would expect trade barriers to be biased away from industries in which the United States has a comparative advantage in international trade because imports in such industries would be expected to be inconsequential and restricting them would provide little benefit.[2] We included a number of variables that previous studies have indicated are associated with U.S. comparative advantage or disadvantage in trade with the simple expectation that trade restrictions should be found predominantly in product lines in which the United States is at a comparative disadvantage vis-à-vis the rest of the world.

Obviously, if there are positive terms of trade effects associated with protection sufficient to offset the deadweight consumption and production costs created by trade barriers, the government can generally increase aggregate social welfare and support for itself by restricting trade. But substantial terms of trade effects are not likely to exist for more than a handful of industries at present levels of protection. In general, increased protection would be expected to promote industry profits at the expense of the general public. In a broad sense, profits should be thought of as the sum of rents to entrepreneurs and to workers with firm- or industry-specific human capital and/or a strong union. The costs to the general public include artificially high product

[1] Assuming that protection is less than prohibitive, that foreign supply depends only on the foreign price, and that the domestic market clearing price, P_D, is positively related to the degree of trade protection given, we can generate the following relationship between industry profits, Π, and trade restrictions:

$$\frac{\partial \Pi}{\partial \tau} = \left\{ S_D + \left(\frac{MC_D}{P_D} - 1 \right) \left[D\eta + S_F \epsilon_F \left(1 - P_F \frac{\partial \tau}{\partial P_D} \right) \right] \right\} \frac{\partial P_D}{\partial \tau}$$

where S_D represents domestic supply, MC_D is the domestic marginal cost of production, D is aggregate domestic demand, η = the absolute value of the own price elasticity of demand for the product, S_F, ϵ_F, and P_F represent the foreign supply, foreign supply elasticity, and foreign price of the product, respectively, and τ represents the tariff or tariff equivalent rate. Details for this derivation are available upon request from the author.

[2] There is, of course, the interesting possibility that the United States has a comparative advantage in some product and that trade restrictions are used to price discriminate between the domestic and foreign market to maximize monopoly rents. That such a phenomenon could be systematic and significant across the 225 manufacturing industries we have data for is doubtful.

prices, misallocation of productive resources, and waste in terms of administrative costs of implementing and maintaining protectionist programs.

In order to rationalize the existence of trade restrictions for numerous commodities in the United States, we assume that, up to some limiting degree of trade restraint, industry interest groups are more responsive to trade policy than consumers are.[3] One simple explanation of such an asymmetry in response to trade policy by industry interests and by consumers stems from the fact that information about market conditions is costly and imperfect and must be updated as market conditions change. To the extent that producers are less numerous than consumers and their well being (wealth) is more directly affected than the wealth of individual consumers by trade policy, they will invest more resources in keeping informed about and be more sensitive to trade policy changes affecting their industry. Clearly, the reverse could be true for industries in which buyers are more concentrated than sellers and the wealth effects are greater per consumer than they are per producer. But such cases are exceptional. This asymmetry in responsiveness does not imply that seller concentration need be positively related to protection. In fact, in his study of the structure of tariffs in Canada, Caves (1976) suggested that concentration could be negatively related to protection. Presumably, the argument would be that the spoils of protection will buy more votes the more widely they are dispersed. In each industry there may exist some critical value of seller concentration relative to buyer concentration, determined by the cost of market information which is required for the industry to obtain protection. And among protected industries there may be a negative association between protection and concentration, CONR. In short, the impact of seller concentration on the existence and extent of trade restrictions is ambiguous.

So far we have argued that producers and workers in an industry have a greater economic incentive on average than consumers in general do to acquire costly information about the impact of trade restrictions on their prices, profits, and wages and in continuing to invest in such information gathering as market conditions change and old information decays or becomes less relevant. However, constant changes in trade restrictions could easily attract media attention and thereby provide consumers with relatively cheap and current infor-

[3] Apart from the observation that nonzero trade restrictions exist for industries that enjoy no obvious terms of trade effects, we provide no direct test of our assumption that industry interests are more responsive than consumer interests to changes in trade policy. However, the notion that industry interests are weighted more heavily than consumer interests in determining the structure of protection is consistent with earlier empirical work by Ray (1974), Baldwin (1976*b*), and Caves (1976).

mation regarding the impact that proposed changes in trade limitations would have on their welfare. Therefore, pressure to alter trade restrictions may follow a fairly discrete time path in order to permit industry interest groups to capitalize on the high cost and low quality of information available to consumers.[4] Maintaining a given quota in the presence of expanding demand or foreign supply would be comparable to imposing higher and higher explicit tariffs over time. Conversely, if demand for a product is declining and/or foreign supply is shrinking, a quota is comparable to a declining explicit tariff. Therefore, assuming some stickiness in policy changes, quotas or other nontariff trade restrictions would have some relative advantage in expanding markets while tariffs would be somewhat advantageous for protecting declining product markets.[5] In our empirical analysis we will attempt to test the relative explanatory power of dynamic demand, ΔD, and supply, ΔS, conditions in tariff and nontariff trade restriction regressions.

Finally, simple demand and supply analysis can be used to demonstrate that the deadweight loss associated with a given trade restriction will be positively related to the absolute value of the own price elasticity of demand for the product, η, and positively related to the domestic elasticity of supply, ϵ_D. To the extent that the government is strongly concerned with the overall welfare effects of trade restrictions, tariff and nontariff trade barriers would be negatively related to the absolute value of the own price elasticity of demand for the product and negatively related to the domestic elasticity of supply for the product.

We can summarize the discussion to this point by indicating the tariff and nontariff trade barrier regressions to be estimated, with the expected sign of the coefficient specified below each variable:

$$\tau = F(\underset{-}{\eta}, \underset{?}{\epsilon_F}, \underset{-}{\epsilon_D}, \underset{-}{X_1}, \ldots, \underset{-}{X_n}, \underset{?}{\text{CONR}}, \underset{+}{\Delta D}, \underset{+}{\Delta S}) \tag{1}$$

and

$$N = G(\underset{-}{\eta}, \underset{?}{\epsilon_F}, \underset{-}{\epsilon_D}, \underset{-}{X_1}, \ldots, \underset{-}{X_n}, \underset{?}{\text{CONR}}, \underset{+}{\Delta D}, \underset{+}{\Delta S}, \underset{+}{\tau}) \tag{2}$$

where N represents a measure of nontariff trade restrictions, the ambiguous sign on the elasticity of foreign supply, ϵ_F, reflects the fact that the sign on the foreign supply elasticity would only be positive in

[4] Similar views regarding the relative merits of tariffs and quotas in the context of dynamic market changes have been expressed by Kreinin (1970), Fishelson and Flatters (1975), Caves and Jones (1977), and others.

[5] Cheh (1974) found some evidence that tariff cuts during the Kennedy Round were less vigorously applied to declining industries in the United States.

the noncompetitive case, ϵ_D represents the domestic supply elasticity, $X_1, \ldots, X_n$ are alternative measures of U.S. comparative advantage, CONR is a measure of seller concentration, and ΔD and ΔS represent the expected percentage growth in demand and supply at the time that trade restrictions are set (both are expected to be more positively significant in explaining nontariff than tariff protection).

In the empirical section we treat the implementation of tariff and nontariff trade barriers as a sequential process. Tariffs predate nontariff barriers in virtually every case, and nontariff trade restrictions are estimated as functions of both tariffs and of the political and economic factors we identify as important explanatory variables. Later we test for this sequential effect and find evidence that it does exist.

II. Empirical Results

The empirical results discussed in this section were generated from data made available by the U.S. International Trade Commission in 1975. More precise definitions of the variables used in the regressions are given in the Appendix. The observations consisted of 225 four-digit manufacturing industries in the United States in 1970 and are specified in a separate appendix available upon request.

Alternative forms for tariff regressions were estimated and are presented in table 1, while estimates of alternative specifications for nontariff trade barrier regressions are presented in tables 2 and 3. For regressions (1.1)–(1.3) in table 1, the dependent variable used was a weighted-average nominal tariff measure for each industry in 1970. In effect, the four-digit tariff rates were calculated by weighting the tariffs of less aggregated components by their import shares within the classification for the given four-digit industry. Regressions (1.4)–(1.6) differ from the earlier ones only in terms of the specification of the dependent variable. In regressions (1.4)–(1.6) the dependent variable is the four-digit industry nominal tariff rate calculated as the simple average of component nominal tariff rates.

The dependent variable in table 2 is an Index of the Incidence of Nontariff Barriers in the U.S., 1970, constructed by the U.S. Tariff Commission.[6] In effect, the index measures the comprehensiveness of

[6] The index was constructed by the U.S. Tariff Commission and is explained in *Trade Barriers*, Report to the Committee on Finance of the U.S. Senate, Part 2, pp. 160–72, Washington, D.C., April 1974. The quantitative restrictions included and the weighting scheme used were as follows: Bilateral Quota (0.91); Global Quota (1.96); Quota (unspecified) (1.36); Prohibited Imports (embargoes) (1.36); State Trading (0.91), Automatic Licensing (0.45); Liberal Licensing (0.45); Discretionary Licensing (0.91); Licensing (unspecified) (0.91); Minimum Price System (1.36); Seasonal Restriction (0.91); Restriction (unspecified) (0.91), Export Restraint (VERs) (1.36); Suspended

U.S. Tariff Regressions, 1970 (OLSQ)†

Independent Variables	Dependent Variables: U.S. Weighted-Average Tariffs 1970			U.S. Simple Average Tariffs: 1970		
	(1.1)	(1.2)	(1.3)	(1.4)	(1.5)	(1.6)
Constant	17.96	16.22	9.94	18.62	17.09	14.95
	(14.15)	(9.74)	(4.32)	(14.14)	(9.76)	(6.05)
FTDNMOS	.	-6×10^{-3}	$.4 \times 10^{-3}$	.	$.3 \times 10^{-3}$	$.9 \times 10^{-3}$
		(.27)	(.21)		(.15)	(.39)
CONR4L	*.01*	*.01*	.02	.02	.02	.02
	(.74)	(.59)	(1.06)	(1.46)	(1.31)	(1.23)
SKILLD	−31.85	−30.96	−33.35	−31.46	−30.78	−32.37
	(9.26)	(8.84)	(10.00)	(8.82)	(8.36)	(9.04)
SEINRD	.	7.95	10.29	. .	5.90	7.32
		(2.13)	(2.86)		(1.51)	(1.90)
ESCAL(67)	−6.69	−6.57	−4.57	−4.11	−4.06	−3.30
	(2.80)	(2.79)	(1.94)	(1.66)	(1.64)	(1.31)
LABINT	.		10.73	. .	. . .	3.36
			(3.63)			(1.06)
KLRA	. . .	-3×10^{-4}	.	. . .	$-.2 \times 10^{-4}$	. . .
		(2.23)			(1.59)	
WDAVPD	.25	−.01	−.18	.14	−.05	−.07
	(.64)	(.04)	(.46)	(.34)	(.11)	(.17)
ΔIMP	.04	.02	$-.45 \times 10^{-2}$	.05	.03	.03
	(.47)	(.25)	(.06)	(.53)	(.39)	(.37)
ΔCON	−.70	−1.11	−.08	−.83	−1.10	−.74
	(.60)	(.96)	(.07)	(.69)	(.90)	(.59)
R^2	.33	.37	.39	.30	.32	.31
F-statistic‡	18.11	13.88	15.29	15.46	11.11	10.88

* Absolute values of *t*-statistics appear in parentheses. For samples of this size *t*-statistics greater than 1.65 are significant at the 5% level and *t*-statistics greater than 2.33 are significant at the 1% level. (For a two-tailed test the values of the *t*-statistic for significance at the 5% and 1% levels, respectively, are 1.96 and 2.576.)

† *OLSQ = Ordinary Least Squares. The weighted average tariff and simple average tariff regression results were not estimated using Tobit since only 5 of the 225 observations were at the limit value of zero for the weighted tariffs and only 1 observation was at the limit value for the simple tariffs.*

‡ For $F(6,218)$ an *F*-statistic greater than 2.80 is in the upper 1% of the distribution. For $F(9,215)$ an *F*-statistic greater than 2.41 is in the upper 1% of the distribution.

TABLE 2

U.S. NONTARIFF TRADE RESTRICTIONS (Tobit)

	Dependent Variable					
	Index of the Incidence of Nontariff Barriers in the U.S.: 1970					
Independent Variables	(2.1)	(2.2)	(2.3)	(2.4)	(2.5)	(2.6)
Constant	1.91	.26	6.68	2.12	.38	6.16
	(1.02)	(.11)	(2.43)	(1.11)	(.16)	(2.13)
FTDNMOS		14×10^{-2}	$.1 \times 10^{-4}$	..	$.12 \times 10^{-2}$	-1×10^{-4}
		(.54)	(.38)		(.46)	(.03)
CONR4L	−.04	−.05	−.06	−.04	−.05	−.06
	(1.92)	(2.36)	(2.86)	(1.98)	(2.44)	(2.74)
SKILLD	−5.05	−6.78	−2.42	−6.26	−8.16	−6.04
	(1.13)	(1.49)	(.53)	(1.42)	(1.82)	(1.37)
SEINRD	.	4.99	2.14	. .	5.95	4.06
		(1.08)	(.47)		(1.29)	(.89)
ESCAL(67)	2.44	2.78	.34	1.61	2.01	−.31
	(.88)	(1.00)	(.12)	(.59)	(.73)	(.11)
LABINT	.		−12.69		. . .	−10.05
			(3.76)			(3.07)
KLRA		1×10^{-5}		.	1×10^{-5}	. .
		(2.14)			(2.02)	
WDAVPD	3.93	−4.004	−3.75	−3.76	−3.87	−3.59
	(3.62)	(3.63)	(3.36)	(3.49)	(3.53)	(3.26)
ΔIMP	−.07	.05	−.03	−.08	−.06	−.05
	(.75)	(.56)	(.34)	(.83)	(.64)	(.50)
ΔCON	1.40	1.34	.09	1.45	1.33	.24
	(1.01)	(.95)	(.07)	(1.04)	(.93)	(.17)
USWTIF	.20	.20	.26	.	. .	. . .
	(2.97)	(2.99)	(3.72)			
USSMPTAR		.	. .	.18	.18	.19
				(2.70)	(2.70)	(2.84)

NOTE.—158 of the 225 observations had the limit value of zero.

nontariff protection in an industry. Fifteen types of nontariff trade restrictions were considered and assigned weights reflecting their relative effectiveness in limiting imports. Then the extent of nontariff protection given to each industry was calculated as the percentage of actual to potential protection to be derived from all 15 nontariff

Import Restriction (0.91), Mixing Regulations (0.91). The weights on the right were determined by assigning each restriction a number from 1 to 3, high or low, depending on its degree of restrictiveness. The average of those numbers was divided into the assigned number for each restriction, resulting in the weights shown. Index values for industry categories were then obtained as follows. For any given commodity, the sum of its specific weights divided by the sum of weights for the 15 categories (11.35) expressed in percentage terms yielded the percentage of maximum nontariff restraint of trade given to that industry. Total imports of major trading countries were used as weights to aggregate the basic product category data. Such a weighting means that restrictions in a heavily traded category were counted more than restrictions in lightly (world) traded categories. In cases in which the listed restrictions were known to apply to only part of a given product category they were arbitrarily counted only half as much as restrictions known to be applicable to an entire category.

TABLE 3

U.S. NONTARIFF TRADE RESTRICTIONS (Probit)

	Dependent Variable					
	Dummy Variable for Nontariff Barriers in the U.S.: 1970					
Independent Variables	(3.1)	(3.2)	(3.3)	(3.4)	(3.5)	(3.6)
Constant	.20	−.26	2.35	.36	−.09	2.36
	(.41)	(.41)	(2.75)	(.72)	(.15)	(2.77)
FTDNMOS	...	$.3 \times 10^{-3}$	$.1 \times 10^{-2}$	..	$-.3 \times 10^{-3}$	1×10^{-2}
		(.34)	(1.20)		(.45)	(1.28)
CONR4L	−.01	−.01	−.02	-7×10^{-2}	−.01	−.02
	(1.38)	(2.36)	(3.10)	(1.46)	(2.41)	(2.96)
SKILLD	−1.79	−2.77	−1.47	−2.13	−3.14	−2.51
	(1.52)	(2.21)	(1.13)	(1.84)	(2.55)	(2.01)
SEINRD	...	2.68	1.86	...	2.84	2.22
		(2.21)	(1.46)		(2.37)	(1.80)
ESCAL(67)	−.82	−.59	−1.73	−.99	−.79	−1.82
	(1.09)	(.74)	(1.96)	(1.34)	(1.00)	(2.14)
LABINT	.	..	−4.79	.	..	−4.01
			(4.75)			(4.22)
KURA	...	$.1 \times 10^{-4}$	...	..	$.1 \times 10^{-4}$	
		(2.95)			(2.78)	
WDAVPD	−.41	−.47	−.40	−.39	−.46	−.38
	(2.53)	(2.78)	(2.22)	(2.39)	(2.68)	(2.16)
ΔIMP	−.27	−2.29	−.02	−.03	−.02	−.02
	(1.05)	(0.87)	(.60)	(1.07)	(.92)	(.69)
ΔCON	.52	.48	.04	.50	.44	.02
	(1.46)	(1.29)	(.11)	(1.41)	(1.19)	(.05)
USWTTF	.06	.06	.08			
	(2.77)	(2.80)	(3.53)			
USSMPTAR		..	..	.05	.05	.05
				(2.49)	(2.51)	(2.77)
Pseudo R^2	.25	.31	.39	.24	.30	.35
Likelihood Ratio Test*	44.29	56.42	73.03	42.29	54.25	65.85

* For eqs. (3.1) and (3.4), a value above 18.50 for the likelihood ratio test is in the upper 1% of the distribution. For the remaining equations, a value above 28.2 for the likelihood ratio test is in the upper 1% of the distribution.

restrictions. Industries with high percentage index values are then viewed as having received more comprehensive nontariff protection than industries with low percentage index values. The dependent variable in table 3 is simply a dummy variable which equals 1 if the index for nontariff barriers is positive and zero otherwise. In every other respect the equations estimated in tables 2 and 3 are the same.

There are three different specifications for each of the tariff relationships and six different specifications for each of the nontariff relationships because we have taken an eclectic approach to the empirical explanation of U.S. comparative advantage. Within each subset of equations the alternative regressions differ only in their specifica-

nontariff trade barriers regressions with a negative significant coefficient.

Caves speculated that a negative relationship between trade restrictions and seller concentration might reflect a bias by politicians to provide protection to industries with many producers in which the benefits would be widely dispersed. In order to test that hypothesis along with the hypothesis that protection has been sequential with nontariff trade restrictions supplementing predetermined tariffs, we ran the simple simultaneous model summarized in table 4. The results presented in table 4 are obtained from two-stage least-squares regression models. The two models differ only in terms of the designated measure for tariff rates. In the first model, we used the weighted-average tariff measure. In the second model, we used the simple average tariff measure. In both models we used the simple 0-1 dummy measure of nontariff barriers to trade. The results obtained in table 4 are qualitatively invariant with the choice of the subset of comparative advantage variables included. Therefore, for the sake of brevity, the estimators presented in table 4 are offered as representative. The principal findings are that tariffs do positively and significantly affect nontariff trade restrictions, while nontariff trade restrictions have no significant impact on tariff determination, supporting our sequential model, and that nontariff trade restrictions are negatively and significantly related to both seller concentration and the geographical concentration of production, GEOG.[9] The negative sign on GEOG means that nontariff trade restrictions are biased toward industries in which production is distributed across regions of the United States in a manner similar to the distribution of population and therefore toward industries with substantial representation in Congress.

We used a measure of scale economies in U.S. manufacturing in 1967, ESCAL(67), as a proxy for the domestic, ϵ_D, elasticity of supply.

[9] GEOG is measured as follows: index value =

$$\sum_{i=1}^{4}\left|\frac{VS_i}{\sum_{i=1}^{4} VS_i} - \frac{P_{OP_i}}{\sum_{i=1}^{4} P_{OP_i}}\right|,$$

where VS_i = value of shipments in region i, $i = 1, \ldots, 4$, and P_{OP_i} = population of region i, $i = 1, \ldots, 4$. To the extent that production in industry i is distributed across the north, south, central, and western United States in a fashion similar to the general population, the index value will approach zero. The basic data for the index are from the *Census of Manufactures* for 1967. The lack of concordance between TCSIC and SIC data means that 27 observations were lost and that table 4 was estimated using 198 rather than 225 observations. Apart from the fact that the inclusion of GEOG in the regression runs for tables 1–3 would have similarly reduced the sample size from 225 to 198, the negative significance of GEOG in the nontariff dummy equations in table 4 indicates that it should have been included in the earlier estimates as well.

TABLE 4*

Simultaneous Estimates of Tariff and Nontariff Trade Restrictions in the U.S., 1970

	Weighted-Average Tariffs		Simple Average Tariffs	
	Dependent Variables			
	U.S. Weighted-Average Tariffs (4.1)	Dummy Variable for Nontariff Barriers (4.2)	U.S Simple Average Tariffs (4.3)	Dummy Variable for Nontariff Barriers (4.4)
Independent Variables	OLSQ	Probit	OLSQ	Probit
Constant	9.69	−2.65	16.02	−7.22
	(2.76)	(1.34)	(4.08)	(2.08)
FTDNMOS	..	$-.18 \times 10^{-2}$		$-.23 \times 10^{-2}$
		(1.61)		(1.96)
CONR4L	.03	−.02	.03	−.02
	(1.97)	(2.81)	(1.41)	(2.91)
SKILLD	−23.52	10.80	−27.02	14.55
	(4.49)	(1.97)	(4.61)	(2.15)
SEINRD		−0.60		0.21
		(0.34)		(0.13)
ESCAL(67)	−4.24	.	−4.83	.
	(1.04)		(1.06)	
LABINT	10.19	−7.42	.79	−.35
	(2.14)	(4.02)	(.15)	(3.41)
KLRA	-2×10^{-4}	1×10^{-4}	$.2 \times 10^{-4}$	13×10^{-4}
	(1.77)	(1.88)	(1.56)	(2.21)
WDAVPD		−.44		−.46
		(2.29)		(2.38)
ΔIMP		$-.9 \times 10^{-2}$		−.29
		(.33)		(1.01)
ΔCON		.55		.81
		(1.17)		(1.56)
PUSWTTF		.52		
		(2.87)		
PUSSMPTAR†	. .	. .		.64
				(2.87)
PDUM2†	.82		.27	
	(.94)		(.28)	
GEOG	1.10	−.92	1.73	−1.53
	(.94)	(2.00)	(1.32)	(2.55)
R^2	.33	. .	.27	
F statistic	13.51		10.25	
Pseudo-R^2	.	.31		.31
Likelihood ratio test	. .	49.10	.	49.10

* Income, using the geographical dispersion index from SIC to TCSIC codes, 27 of the 225 observations had to be deleted for missing data. Consequently, the models estimated in table 4 are based on 198 observations.

† PUSWTTF, PUSSMPTAR, and PDUM2 are the predicted values of USWTTF, USSMPTAR and DUM2 on the instrumental variables from the first stage of the two-stage least-squares estimation procedures.

The discussion of Section I suggested that ESCAL(67) should be negatively and significantly related to tariff and nontariff trade restrictions if the government is strongly concerned about the overall welfare costs of protection. ESCAL(67) appears in all of the weighted tariff and nontariff trade barrier dummy regressions with the expected negative sign and generally significant coefficient. While the sign of ESCAL(67) is negative, it is less significant in the simple tariff regressions.

In Section I we indicated that we expected to find a negative and significant coefficient for the absolute value of the own price elasticity of demand for the product in the trade restriction equations. On the assumption that the own price elasticity of demand is reduced by product heterogeneity, we used a measure of product differentiation, WDAVPD, as an inverse measure of the absolute value of the own price elasticity of demand. Consequently, we expected to find a positive and significant coefficient on the product differentiation variable in the tariff and nontariff trade barrier regressions. The relative negative significance of WDAVPD in the nontariff trade barrier regressions compared with the tariff regressions may be partially related to another factor that we have not discussed. Clearly, nonprice, quantitative restrictions are more easily applied and enforced if products are fairly standardized. Consequently, the negative coefficient on product heterogeneity may partially reflect an administrative preference by the government for nonprice restrictions for homogeneous products.

Earlier we argued that while market conditions change continuously, it may still make sense for industry interest groups to push for changes in trade restrictions in discrete fashion. If so, we would expect tariffs to be relatively more productive in industries that are expected to contract and nontariff trade barriers to be relatively more productive in expanding industries. We used the percentage growth in apparent domestic consumption between 1965 and 1970, measured by the percentage growth in domestic shipments plus imports minus exports, as a proxy for demand shifts, ΔD, and the percentage growth in imports between 1965 and 1970 as a proxy for supply shifts, ΔS. Clearly, both the apparent domestic consumption variable and the import expansion variable reflect both demand and supply phenomena and in that sense are poor proxies for the effects of pure demand and supply shifts that we are trying to measure. With that caveat in mind, we simply report that neither of the variables included to measure dynamic influences has any significant explanatory power in the tariff or nontariff trade barrier regressions.

Unfortunately, we have no good measures of the industry-specific revenue effects associated with tariff and nontariff trade restrictions.

or of the administrative costs of implementing alternative forms of trade restrictions across industries. Consequently, all of our estimates suffer from the potential biases associated with important left-out variables.

III. Conclusions

We presented evidence that both tariff and nontariff trade restrictions are found predominantly in industries in which the United States has no comparative advantage vis-à-vis the rest of the world and away from industries in which the deadweight losses to consumers from protection would be high. In addition, we presented direct evidence that nontariff trade restrictions have been used to supplement tariffs and, thereby, offset the trade-liberalizing effects of post–World War II tariff agreements.

We also presented evidence that there are significant differences in the industrial characteristics of industries with tariff protection compared with those with nontariff trade protection. Specifically, tariffs are positively and significantly related to labor intensity and negatively and significantly related to the capital/labor ratio while just the opposite is true for nontariff trade restrictions. In addition, nontariff trade restrictions are negatively and significantly related to both seller concentration and geographical concentration in an industry while both characteristics are positively and insignificantly related to tariffs.

Appendix

Definitions of Variables

Independent Variables

FIDNMOS—Product cycle proxy: unweighted average of Schedule B first trade dates corresponding to TCSIC as of January 1974.

CONR4L —Concentration Ratio, 1970: percentage of shipments accounted for by the four largest firms in the industry.

SKILLD —Skills measure, 1970: professional and kindred workers, plus managers and administrators (except farm), plus craftsmen and kindred workers, as a percentage of total employment. Based on three-digit SIC data with values repeated at four-digit levels.

SEINRD —Percentage of scientists and engineers in R & D in 1970. Based on two-digit SIC data with values repeated at the four-digit level.

SCAL(67)—Economies of Scale measure, 1970: value of the exponent in the regression equation $V = KN^{a}$, where V is the ratio of value added in plants employing N persons to average values added for the industry, and K is a constant.

LABINT —Labor Intensity Ratio, 1970 (payroll divided by value added).
KLRA —Capital/Labor Ratio, 1970: Total capital stock divided by employment (thousands of dollars).
WDAVPD —Weighted Average of Product Differentiation measure, 1970: The measure is the coefficient of variation in the unit values of exports destined to different countries weighted by U.S. export shares, i.e., the standard deviation of U.S. export unit values weighted by export shares divided by means of unit values.
ΔIMP —Percentage change in imports between 1965 and 1970.
ΔCON —Percentage change in apparent U.S. consumption between 1965 and 1970, where apparent U.S. consumption is measured by the value of U.S. shipments plus imports minus exports.
GEOG —See n. 9

Dependent Variables

Tariffs—USWTTF—U.S. weighted-average tariffs for 1970 and USSMPTAR—U.S. simple average tariffs for 1970 are taken from the GATT *Tariff Study*, Geneva, 1970.

Index of Nontariff Barriers—Index of the Incidence of Nontariff Barriers in the U.S., 1970. The index was constructed by the U.S. Tariff Commission and is explained in *Trade Barriers*, Report to the Committee on Finance of the U.S. Senate, Part 2, pp. 160–72, Washington, D.C., April 1974. See n. 6.

Dummy Variable for Nontariff Barriers in the U.S., 1970—DUM2—The value of the dummy variable was set equal to 1 if the index of nontariff barriers was nonzero for a particular product and zero otherwise.

References

Baldwin, Robert E. "Determinants of the Commodity Structure of U.S. Trade." *A.E.R.* 61 (March 1971): 126–46.

——— "Reply." 62 (June 1972): 465.

——— "Trade and Employment Effects in the United States of Multilateral Tariff Reductions." *A.E.R. Papers and Proc.* 66 (May 1976): 142–48. (*a*)

——— "The Political Economy of Postwar U.S. Trade Policy." *Bulletin*, New York Univ. Graduate School of Business, no. 4 (1976). (*b*)

Baldwin, Robert E., and Lage, Gerald M. "A Multilateral Model of Trade-balancing Tariff Concessions." *Rev. Econ. and Stats.* 53 (August 1971): 237–44.

Caves, Richard E. "Economic Models of Political Choice: Canada's Tariff Structure." *Canadian J. Econ.* 9 (May 1976): 278–300.

Caves, Richard E., and Jones, Ronald W. "Tariff Policy and Trade Liberalization." *World Trade and Payments: An Introduction*. 2d ed. Boston: Little, Brown, 1977.

Cheh, John H. "United States Concessions in the Kennedy Round and Short-Run Labor Adjustment Costs." *J. Internat. Econ.* 4 (November 1974): 323–40.

———. "A Note on Tariffs, Nontariff Barriers, and Labor Protection in United States Manufacturing Industries." *J.P.E.* 84, no. 2 (April 1976): 389–94.

Fishelson, Gideon, and Flatters, Frank. "The (Non)Equivalence of Optimal

Tariffs and Quotas under Uncertainty." *J. Internat. Econ.* 5 (November 1975): 385–93.

Helleiner, G. K. "The Political Economy of Canada's Tariff Structure: An Alternative Model." *Canadian J. Econ.* 10 (May 1977): 318–26.

Hufbauer, Gary C. "The Impact of National Characteristics and Technology on the Commodity Composition of Trade in Manufactured Goods." In *The Technology Factor in International Trade*, edited by Raymond Vernon. New York: Columbia Univ. Press (for Nat. Bur. Econ. Res.), 1970.

Kreinin, Mordechai E. "More on the Equivalence of Tariffs and Quotas." *Kyklos* 23, fasc. 1 (1970): 75–79.

Pincus, J. J. "Pressure Groups and the Pattern of Tariffs." *J.P.E.* 83, no. 4 (August 1975): 757–78.

Ray, Edward John. "The Optimum Commodity Tariff and Tariff Rates in Developed and Less Developed Countries." *Rev. Econ. and Stats.* 56 (August 1974): 369–77.

Stern, Robert M. "The U.S. Tariff and the Efficiency of the U.S. Economy." *A.E.R. Papers and Proc.* 54 (May 1964): 459–70.

Stone, Joe A. "A Comment on Tariffs, Nontariff Barriers, and Labor Protection in United States Manufacturing Industries." *J.P.E.* 86, no. 5 (October 1978): 959–62.

Inflation, Corporate Income Taxation, and the Demand for Capital Assets

Richard W. Kopcke

Federal Reserve Bank of Boston

The demand for capital is not systematically related to either the level or the rate of change of "effective" income tax rates on corporate capital assets. Rising inflation during the last 10 years has raised the user cost of capital for durable assets relative to that for short-lived assets even though this inflation has raised effective tax rates for more durable capital less than for short-lived assets. Even with replacement-cost depreciation allowances, the level and pattern of investment incentives probably will continue to vary with the inflation rate.

The past decade of high and rising inflation rates has encouraged the study of U.S. income tax codes and inflation's influence on the demand for capital assets. Many have examined the "effective income tax rate" on capital to describe how inflation has altered both the level and pattern of investment incentives. In a recent article, for example, Auerbach (1979) observes: "With a positive rate of inflation, a historic cost depreciation rule biases the choice of asset life toward greater durability" (p. 621). Such a conclusion, based on the behavior of effective income tax rates, contradicts the spirit, if not the letter, of much traditional analysis of investment incentives and the demand for capital (see, e.g., Black 1959; Brown 1962; Samuelson 1964; Hall and Jorgenson 1967; Musgrave and Musgrave 1976; and Boadway 1978) so it warrants examination.

One principal conclusion of this paper is that the demand for capital is not systematically related either to the level or to the rate of

The analysis and conclusions are not necessarily endorsed by the Federal Reserve Bank of Boston or the Federal Reserve System.

[*Journal of Political Economy*, 1981, vol. 89, no. 1]
© 1981 by The University of Chicago. 0022-3808/81/8901-0005$01.50

change of the effective corporate income tax rates on capital assets. Another principal conclusion is that the rising domestic inflation rate of the past 10 years not only has reduced investment incentives generally, but it has reduced the incentive to acquire more durable capital goods most severely. Even if U.S. corporate income tax codes were amended to embrace replacement-cost depreciation allowances, investment incentives would not necessarily be independent of asset service lives nor would the relative demand prices of capital goods necessarily be independent of the inflation rate.

I. Inflation and the Demand for Capital

Assuming perfect markets and certainty of foresight, investors maximize their wealth by purchasing capital yielding no less than ρ, the real discount rate for after-tax corporate cash flow. In other words, the real price of capital, P (the output price is the numeraire),[1] equals

$$\int_0^\infty e^{-\rho t}[(1-\tau)e^{-\delta t}Q_k + \tau PD(t,\delta)e^{-\Pi t}]dt + P \cdot ITC, \text{ or} \qquad (1)$$
$$P = \frac{(1-\tau)}{(\rho+\delta)}Q_k + \tau PZ + P \cdot ITC,$$

where τ is the corporate tax rate, δ is the rate of decay of capital, Q_k is the marginal physical product of a unit of new capital, $D(t, \delta)$ is the schedule of depreciation allowances per dollar of investment undertaken t years ago, ITC is the rate of investment tax credit, Z is the present value of depreciation allowances per dollar of investment, and Π is the inflation rate. Formula (1) represents the present value of after-tax operating income plus the present value of the tax shelter afforded by depreciation allowances and tax credits; $D(t, \delta)$ is "discounted" by Π because depreciation allowances are linked to the original purchase price of capital assets, so the real value of these allowances must decline as nominal capital goods prices rise.

In equilibrium, then,

$$CS = P_s(\rho + \delta_s)(1 - \tau Z_s)/(1-\tau) = Q_{ks},$$
$$CE = P_e(\rho + \delta_e)(1 - \tau Z_e - ITC)/(1-\tau) = Q_{ke}, \qquad (2)$$

or the user cost of capital (Hall and Jorgenson 1967)—CS for nonresidential structures, CE for equipment—equals the marginal physical product of capital. Because structures and equipment are each

[1] For the most part, the notation matches that in Auerbach's (1979) article. P may equal unity as it does in Auerbach's article; however, it is often useful to relax this assumption because the relative prices of capital assets are not constant.

represented by a variety of capital goods bearing different rates of decay while offering different flows of capital services, $A(\delta)Q_k$, equilibrium for both equipment and structures demand prevails when

$$\left[\frac{CS(\delta_s)}{A(\delta_s)}\right]\Big/\left[\frac{CE(\delta_e)}{A(\delta_e)}\right] = Q_{ks}/Q_{ke}. \tag{3}$$

Referring to the equilibrium condition described in expression (2), it is evident that a changing inflation rate, Π, can alter the demand for capital. Higher inflation rates, other things equal, depress the present value of real depreciation allowances; this, in turn, raises the user cost of capital. From (3), it is also evident that if higher inflation rates do not increase the user costs of all capital goods proportionately, the pattern of investment demand will change as well.

In equilibrium,

$$\frac{CS(\delta_1)}{CS(\delta_2)} = \frac{A(\delta_1)}{A(\delta_2)}. \tag{4}$$

Should changing tax rates tend to increase $CS(\delta_1)$ relative to $CS(\delta_2)$, $P_s(\delta_1)$ and $P_s(\delta_2)$ will compensate so that equilibrium is maintained. Of course, if the new balance requires the price of more durable structures, $P_s(\delta_1)$, to decline relative to that of less durable structures, $P_s(\delta_2)$, and if the supply of structures varies with price, then the stock of nonresidential structures will become less durable. The influence of inflation on the demand price of capital and the choice of asset life depends on the sensitivity of the present value of depreciation allowances to changing inflation rates, $\partial Z/\partial \Pi$, as well as the importance of these allowances as a tax shelter, $(1 - \tau Z - ITC)$.

Table 1 shows how rising inflation rates reduce the present value of real depreciation allowances for a variety of capital goods according to current U.S. corporation tax law. The table illustrates two important points. First, the value of depreciation allowances is lower for longer-lived capital assets. Second, with higher inflation, the value of depreciation allowances drops at a faster rate for more durable capital. Therefore, the modest role of depreciation allowances in the cash flow accruing to durable capital does not mean that these assets are necessarily less sensitive to changing inflation rates than are less durable assets.

In fact, as shown in table 2, the tax shelter offered by depreciation allowances is sufficiently important for long-lived capital that higher inflation rates increase the user cost of capital for these assets more than that of less durable assets. As the annual inflation rate rises from zero to 6 percent, other things equal, the investment incentives for 40-year structures fall 22 percent, but the incentives for 20-year structures fall only 19 percent. Investment incentives generally drop much less for the tabulated categories of equipment.

TABLE 1

THE PRESENT VALUE OF REAL DEPRECIATION ALLOWANCES PER DOLLAR OF INVESTMENT

A. PRODUCERS' DURABLE EQUIPMENT*

Inflation Rate (%)	Asset Lifetime (Years) 5	10	15
0	.44	43	.42
2	.42	.40	.38
4	.41	37	34
6	.39	.35	31
8	.37	.33	29
10	36	31	27
20	.30	.23	.19
30	.25	18	14

B. NONRESIDENTIAL STRUCTURES†

Inflation Rate (%)	Asset Lifetime (Years) 20	30	40
0	.41	38	36
2	35	31	.28
4	30	.26	.22
6	.27	.22	19
8	24	19	.16
10	22	17	14
20	15	10	.09
30	11	.08	06

NOTE.—Entries equal nominal depreciation allowances multiplied by .46 the marginal corporate income tax rate discounted by (1.015)(1 + Π), where 1.5 percent is the real after-tax risk-adjusted discount rate (see Brown 1962) and Π is the inflation rate.

* Nominal depreciation allowances are calculated using the sum-of-the-years'-digits formula.

† Nominal depreciation allowances are calculated using the 150 percent–declining formula with a switch to straight-line allowances after one-third of the asset's life has passed.

Because rising inflation reduces the present value of depreciation allowances most quickly for the longest-lived assets, however, the user cost of capital for these assets becomes progressively less sensitive to further increases in the inflation rate. For example, if the inflation rate were to rise from 6 to 10 percent, investment incentives would drop another 5 percent for 40-year structures compared with an incremental decline of 6 percent for 20-year structures. Should the inflation rate exceed 20 percent, the erosion of real depreciation allowances would be so severe for the shorter-lived structures that longer-lived structures become relatively more attractive investments.

Despite a popular contention that the value of the depreciation tax shelter, hence its annual loss in value due to inflation, is greatest for

TABLE 2

INFLATION AND INVESTMENT INCENTIVES

A. PRODUCERS' DURABLE EQUIPMENT

INFLATION RATE (%)	ASSET LIFETIME (Years)		
	5	10	15
0	1.00	1.00	1.00
2	.96	.94	.92
4	.93	.89	.86
6	.90	.85	.81
8	.87	.81	.78
10	.85	.78	.75
20	.77	.70	.67
30	.72	.65	.63

B. NONRESIDENTIAL STRUCTURES

INFLATION RATE (%)	ASSET LIFETIME (Years)		
	20	30	40
0	1.00	1.00	1.00
2	.91	.89	.88
4	.85	.83	.82
6	.81	.79	.78
8	.78	.76	.76
10	.76	.74	.74
20	.69	.69	.70
30	.67	.67	.68

NOTE.—Entries for each asset equal the reciprocal of the user costs of capital, expressions (2), divided by the reciprocal of user costs when the inflation rate equals zero. Values of Z_t and Z_s are taken from table 1, and, for equipment, ITC equals .1 except for 5-year assets for which ITC equals .067. The figures in this table are independent of the pattern of "true" capital decay—inflation alters the relative user cost through the factor $(1 - \tau Z - ITC)$.

less durable assets, it is the percentage change in the effective purchase price of capital, $P(1 - \tau Z - ITC)$, that influences investment incentives. Table 2 shows that higher inflation rates raise this effective purchase price most for more durable, not less durable, assets for relevant inflation rates in the United States. The more durable the asset, the more its depreciation allowances are postponed, hence the more inflation may "discount" the value of these allowances before they may be claimed, as shown in the first table. Consequently, high inflation rates discourage the purchase of longer-lived capital more than short-lived assets.[2] Ironically, the class of investment goods most

[2] This is true whenever the pattern of investment demand for high inflation rates is compared with the pattern of demand for inflation rates near zero. As explained in the text, however, as the inflation rate rises from an already high level, the relative demand price of some longer-lived assets may rise.

needing "protection" from the high inflation rates of the past 15 years—nonresidential structures—has not received the benefit of the tax credits and accelerated depreciation granted to producers' durable equipment.

This shift in the relative capital costs that inflation induces will alter investment demand to favor less durable capital goods, thereby depressing the relative demand price for more durable capital assets. As a result, the stock of more durable structures, other things equal, will fall relative to that of less durable structures. To the extent that structures and equipment are substitutes, the stock of structures generally will decline relative to that of equipment. Of course, the higher user cost for all capital goods implies a declining capital-labor ratio in addition to this substitution among varieties of capital assets.

Assuming that the real price for all capital goods equals unity, a unit of output may be consumed or installed either as short-lived equipment or as a very durable structure; consequently, changing rates of inflation cannot affect asset prices. Rather, from expression (4), if $CS(\delta)/A(\delta)$ for a more durable factory exceeds that for a less durable factory due to increasing inflation, investors will install less durable structures; more durable factories are now economically obsolete.

In summary, with current U.S. income tax codes, which link depreciation allowances to the original purchase price of capital assets, changing inflation rates alter the relative user costs of capital, other things equal. Higher inflation rates initially raise these user costs more for long-lived assets than for short-lived assets, thereby reducing the average service life of the capital stock.

II. Inflation and the Effective Corporate Tax Rate

Auerbach (1979) defines the investor's implicit discount rate as

$$v \equiv c - \delta, \tag{5}$$

and then the effective corporate tax rate,

$$\theta \equiv (v - \rho)/v. \tag{6}$$

In view of expressions (2) and (3), it is possible for θ either to rise or to fall with the service lives of capital assets. Consider, however, the simple case wherein $A(\delta)$ equals unity and prices, $P(\delta)$, are flexible for all assets (the definition of a unit of capital has eliminated the apparent differences in productivity among pure substitutes); then (from [4]), $CS(\delta_1) = CS(\delta_2)$ and $\theta(\delta_1) > \theta(\delta_2)$: The effective tax rate on the more durable factory is largest, though the user costs are equal.

As explained in the previous section, a higher inflation rate will tend to raise the user cost for long-lived structures more than that of less durable structures. A higher inflation rate, then, will also increase

Q_{ks} by reducing the stock of nonresidential structures, other things equal. Retaining the assumption that $A(\delta)$ equals unity, from (4) $CS(\delta_1)$ will still equal $CS(\delta_2)$ in equilibrium, and both will be greater, with higher inflation rates as long as capital goods prices are flexible. Though $CS(\delta_1)$ and $CS(\delta_2)$ rise proportionately, $\theta(\delta_1)$ increases less than $\theta(\delta_2)$ because δ_1, δ_2, and ρ do not change. Even though the effective tax rate, θ, on less durable capital rises more than that on longer-lived assets, the capital stock is comprised of more short-lived assets with higher inflation rates because $P_k(\delta_1)$ (the price of the more durable asset) has fallen relative to $P_k(\delta_2)$ in equilibrium, as described after expression (4).

If $A(\delta)$ rises with δ and if the prices of all capital were unity, a medium-lived asset may have a relatively high user cost and yet be the asset in greatest demand due to its high productivity, $A(\delta)$. Consequently, the effective tax rate, θ, for this asset may exceed that for many alternative assets not favored by investors.

Assume, for $\delta_1 < \delta_2 < \delta_3$, that

$$\frac{CS(\delta_1)}{A(\delta_1)} > \frac{CS(\delta_2)}{A(\delta_2)} < \frac{CS(\delta_3)}{A(\delta_3)}, \text{ and} \qquad (7)$$

$$CS(\delta_1) < CS(\delta_2) < CS(\delta_3).^3$$

It is reasonable to assume $CS'(\delta) < 1$ (see n. 3); therefore,

$$\begin{aligned} \theta(\delta_1) &= [CS(\delta_1) - \delta_1 - \rho]/[CS(\delta_1) - \delta_1] > \\ \theta(\delta_2) &= [CS(\delta_2) - \delta_2 - \rho]/[CS(\delta_2) - \delta_2] > \\ \theta(\delta_3) &= [CS(\delta_3) - \delta_3 - \rho]/[CS(\delta_3) - \delta_3]. \end{aligned} \qquad (8)$$

The medium-lived asset is the optimal investment according to (7), but expression (8) reveals that its effective tax rate is not the lowest.

Just as the level of θ is a poor guide to investment incentives, so the change in the effective tax rate is a poor guide to the change in investment incentives. From (7), the short-lived asset (δ_3) will be preferred to the medium-lived asset once inflation is high enough to cause

$$\frac{CS(\delta_3)}{CS(\delta_2)} < \frac{A(\delta_3)}{A(\delta_2)}, \qquad (9)$$

because $\partial[CS(\delta_3)/CS(\delta_2)]/\partial\Pi < 0$.[4]

[3] $CS'(\delta) = [(1 - \tau Z) - \tau(\rho + \delta)Z'(\delta)]/(1 - \tau)$. According to table 1, $\tau Z'(\delta) \approx 2$, so the first term in the brackets is approximately 0.6 and the second is approximately 0.3; therefore, it is reasonable to assume $0 < CS'(\delta) < 1$. Similarly, $CS''(\delta) < 0$.

[4] As shown in table 2, unless inflation rates exceed 20 percent, a higher inflation rate raises the user cost for less durable structures (δ_3) proportionately less than the user cost of more durable structures (δ_2). The relative user cost of less durable structures declines as inflation rises.

Though rising inflation depresses the user cost of short-lived assets relative to that of longer-lived assets, however, it need not depress the *relative effective tax rate on shorter-lived assets: $\partial\theta(\delta_3)/\partial\Pi$ is not necessarily less than $\partial\theta(\delta_2)/\partial\Pi$, nor is $\partial[\theta(\delta_3)/\theta(\delta_2)]/\partial\Pi < 0$ always. From expressions (7) and (8), for example,*

$$\left[\frac{\partial\theta(\delta_3)}{\partial\Pi}\right]\Big/\left[\frac{\partial\theta(\delta_2)}{\partial\Pi}\right] = \frac{\partial \ln [CS(\delta_3)]/\partial\Pi}{\partial \ln [CS(\delta_2)]/\partial\Pi} \cdot \frac{CS(\delta_3)}{CS(\delta_2)} \cdot \frac{[CS(\delta_2) - \delta_2]^2}{[CS(\delta_3) - \delta_3]^2}. \quad (10)$$

The product of the last two factors in the right-hand side of (10) exceeds unity;[5] consequently, the left-hand side of (10) may exceed unity—the tax rate on less durable assets may rise more than that on longer-lived assets—even though $CS(\delta_3)/CS(\delta_2)$ is falling. Whenever $\theta(\delta_3)$ rises more than $\theta(\delta_2)$, $\theta(\delta_3)/\theta(\delta_2)$ also increases because $\theta(\delta_3) < \theta(\delta_2)$ in this example. While relative user costs for shorter-lived structures fall with rising inflation, the effective tax rate on these assets may rise more than that on longer-lived structures.

Finally, from (2), (5), and (6), assuming $P \equiv 1$,

$$\theta = [\tau(1 - Z) - ITC]/[1 - \tau Z - ITC - (1 - \tau)\delta/(\rho + \delta)]. \quad (11)$$

Assuming that the intertemporal discount rate (ρ) is constant, the effective tax rate on corporate capital must vary with the inflation rate unless the statutory tax rate (τ) equals zero or depreciation allowances are adjusted for changing capital goods prices—$D(t, \delta)$ is replaced by $D(t, \delta)e^{\Pi t}$ in (1)—so that $\partial Z/\partial\Pi = 0$; $D(t, \delta)$ may correspond to any capital consumption scheme, including "economic depreciation."

It is likely, however, that ρ changes whenever the inflation rate changes;[6] therefore, the demand for capital may change whenever Π

[5] See n 3. Because $CS'(\delta) > 0$, $CS(\delta_3) > CS(\delta_2)$. $CS'(\delta) < 1$, however, implies that $[CS(\delta_2) - \delta_2]^2 > [CS(\delta_3) - \delta_3]^2$. Therefore, both of the last two factors of (10) exceed unity.

[6] In many standard neoclassical growth models, combining intertemporal consumption and production decisions (e.g., Sidrauski 1967; Intriligator 1971), ρ by hypothesis is constant. In other versions of this model the intertemporal utility function is replaced by a saving rate and portfolio balance relationship (Tobin 1955, 1965) In these latter models, higher inflation rates can lower the real return on money balances, thereby encouraging investors to acquire more capital, by reducing ρ, until its marginal physical product drops enough to restore portfolio equilibrium. In other models, ρ is neither constant nor essentially equal to the real return on money. Inflation is a by-product of public and private policies intended to redistribute resources (Schumpeter 1939, Robinson 1954, 1962; Pasinetti 1962). This reallocation of wealth, income, or spending undoubtedly will change the social rate of time preference, ρ. Although Tobin's growth model suggests ρ would fall as inflation increases, Schumpeter's model suggests ρ would rise It is not unreasonable, then, to assume that higher inflation depresses ρ and, in turn, increases the average service life of capital if corporate income taxes are indexed. Yet, so little is known about the social rate of time preference that it is unreasonable to ignore the alternative conclusion: Higher inflation rates accompany an increase in ρ, thereby depressing the average service life of capital.

changes, unless this demand is also independent of ρ. By setting $ITC = 0$ and either setting $\tau = 0$ or allowing depreciation schedules that are a linear combination of "economic depreciation" adjusted for changing capital goods prices and immediate expensing of investment expenditures, not only is the effective tax rate independent of the inflation rate, but it is independent of the discount rate, ρ, and asset service lives. (Even if depreciation allowances were "indexed" for inflation, unless these allowances correspond to the pattern of an asset's "economic depreciation," the effective tax rate will not be independent of the asset's service life [see Samuelson 1964].) Of course, these last reforms accomplish this goal by pegging the effective tax rate at zero. Nevertheless, even though $\theta = 0$ for all values of Π, δ, and ρ, these reforms cannot make the user cost of capital independent of ρ and δ because (from [2]):

$$C = (\rho + \delta). \tag{12}$$

The demand for capital, therefore, is not systematically related either to the level or to the rate of change of the effective corporate tax rate defined by expressions (5) and (6). The demand for capital depends on the user cost of capital and the marginal physical product of capital alone in neoclassical macro models, so the income tax codes influence investment incentives only by altering capital's user cost. For studying the demand for capital, then, the suggested effective corporate tax rate, θ, is not a sufficient statistic—it alone is not a definitive measure of investment incentives. Conversely, the user cost of capital does not measure the effective tax burdens borne by the various capital assets. Each concept—the cost of capital and the effective tax rate—serves a particular purpose, and their roles are distinct.

III. Summary

Unless the aggregate rate of time preference has declined substantially, the rising inflation rate of the past 10 years not only has reduced business's demand price for capital generally but it has reduced the relative demand price for more durable assets most severely. Even if the corporate tax rate were zero or depreciation allowances were a linear combination of economic depreciation adjusted for changing capital goods prices and immediate expensing of investment expenditures, thereby pegging the effective corporate income tax rate at zero, the level and pattern of investment incentives most likely would vary with the inflation rate.

Of course, it is one thing for inflation to influence investment incentives directly; it is quite another for changing investment incentives to accompany a changing inflation rate because the intertem-

poral discount rate varies with the inflation rate. Although it is probably unreasonable to expect any tax reform to insulate the demand price of capital from the influence of policies that reallocate resources, any schedule of depreciation allowances adjusted for changing prices can insulate investment incentives from inflation per se.

References

Auerbach, Alan J. "Inflation and the Choice of Asset Life." *J.P.E.* 87, no. 3 (June 1979): 621–38.

Black, John. "Investment Allowances, Initial Allowances and Cheap Loans as Means of Encouraging Investment." *Rev. Econ. Studies* 27 (October 1959): 44–49.

Boadway, Robin W. "Investment Incentives, Corporate Taxation and Efficiency in the Allocation of Capital." *Econ. J.* 88 (September 1978): 470–81.

Brown, Edgar C. "Tax Incentives for Investment." *A.E.R. Papers and Proc.* 52 (May 1962): 335–45.

Hall, Robert E., and Jorgenson, Dale W. "Tax Policy and Investment Behavior." *A.E.R.* 57 (June 1967): 391–414.

Intriligator, Michael D. *Mathematical Optimization and Economic Theory.* Englewood Cliffs, N.J.: Prentice-Hall, 1971.

Musgrave, Richard A., and Musgrave, Peggy B. *Public Finance in Theory and Practice.* 2d ed. New York: McGraw-Hill, 1976.

Pasinetti, Luigi L. "Rate of Profit and Income Distribution in Relation to the Rate of Economic Growth." *Rev. Econ. Studies* 29 (October 1962): 267–79.

Robinson, Joan. "The Production Function and the Theory of Capital." *Rev. Econ. Studies* 21, no. 2 (1954): 81–106.

———. *Essays in the Theory of Economic Growth.* New York: St. Martin's, 1962.

Samuelson, Paul A. "Tax Deductibility of Economic Depreciation to Insure Invariant Valuations." *J.P.E.* 72, no. 6 (December 1964): 604–6.

Schumpeter, Joseph A. *Business Cycles.* New York: McGraw-Hill, 1939.

Sidrauski, Miguel. "Rational Choice and Patterns of Growth in a Monetary Economy." *A.E.R. Papers and Proc.* 57 (May 1967): 534–44.

Tobin, James. "A Dynamic Aggregative Model." *J.P.E.* 63, no. 2 (April 1955): 103–15.

———. "Money and Economic Growth." *Econometrica* 33 (October 1965): 671–84.

A Monetary Approach to the Crawling-Peg System: Theory and Evidence

Mario I. Blejer

Hebrew University of Jerusalem

Leonardo Leiderman

Tel-Aviv University

This paper develops and estimates a model of the joint determination of the exchange rate, international reserves, and the rate of inflation under a crawling-peg system. The framework presented, which is an extension of previous work on the monetary approach, generates short-run deviations from purchasing power parity that occur simultaneously with movements in both international reserves and the exchange rate. The model is estimated by full-information maximum likelihood on the basis of quarterly data for Brazil.

With the advent of the modern literature on the monetary approach to the balance of payments and to the exchange rate, and following the seminal contributions of Mundell (1968, 1971), Johnson (1972), and Dornbusch (1973), a large amount of effort has been devoted to emphasizing the role of the money market and asset markets in the determination of the balance of payments and the exchange rate. The

We are also research associates at the Center for Latin American Development Studies of Boston University, whose support in the preparation of this paper is gratefully acknowledged. We are indebted to D. Cavallo, R. Dornbusch, J. A. Frenkel, G. Hanoch, A. Harberger, R. Hodrick, A. Stockman, and referees of this *Journal* for valuable comments. Previous versions of this paper were presented at the P. Sapir Conference on Inflation and Development, Tel Aviv; the Summer Meetings of the Econometric Society, Montreal, and a seminar at the Fundacion Mediterranea, Córdoba, Argentina.

[*Journal of Political Economy*, 1981, vol. 89, no. 1]
© 1981 by The University of Chicago. 0022-3808/81/8901-0002$01.50

long-run effects of monetary shocks in a regime of fixed exchange rates, the effects of devaluations and changes in commercial policy in a monetary setting, and the short-run consequences of monetary imbalance on the inflation rate and the balance of payments have recently been the object of careful theoretical and empirical study. The importance of monetary variables in the determination of the exchange rate in a regime of floating rates has also been analyzed in detail.[1]

Most of these studies, however, have dealt either with cases in which the exchange rate has been kept fixed for very long periods or with free-floating cases in which the exchange rate is predominantly determined by the interaction of market forces without government intervention.[2]

Yet in recent years many countries have experienced simultaneous fluctuations in their exchange rates and international reserves. These fluctuations have occurred either under a managed-float system, characterized by government intervention in the foreign-exchange market, or under a crawling-peg system, in which the authorities determine, and periodically change, the country's exchange rate.[3] In order to analyze the experience of countries with these characteristics, an extension to the previous formulations of the monetary approach is required.

In this paper we develop and estimate a model for the analysis of the joint determination of the exchange rate, international reserves, and the rate of inflation under a crawling-peg system. Section I

[1] A number of important contributions on the topic are collected in Frenkel and Johnson (1976, 1978) and in the International Monetary Fund (1977) volume. A review of the empirical evidence on the monetary approach to the balance of payments and exchange rates is presented in Magee (1976). Johnson (1977) and Whitman (1975) present comprehensive reviews of the recent contributions to the monetary-approach literature, while a critique of some of the central aspects of that approach is found in Hahn (1977). All these deal with fixed-exchange-rate regimes. Among those dealing with floating rates are Kouri (1975), Dornbusch (1976), Frenkel (1976), Mussa (1976), and Frenkel and Clements (in press).

[2] Exceptions are the studies by Artus (1976) and by Girton and Roper (1977). The latter develop and test a monetary model of exchange-market pressure, which is defined as a composite variable that includes changes in both the exchange rate and international reserves. This model was recently applied to Brazil by Connolly and da Silveira (1979) who, using annual data, attempt to explain the behavior of the exchange-market pressure variable. As will be seen, our model is designed to deal simultaneously with the joint determination of each of the two components of exchange-market pressure as well as of the domestic rate of inflation.

[3] A theoretical analysis which concludes that the optimum exchange-rate regime will correspond to neither of the extremes of a completely fixed or a completely flexible rate is presented by Boyer (1978) and Frenkel (1980), who concludes that "for the purpose of empirical work it is useful to design a framework of the adjustment mechanism that can accommodate simultaneously changes in international reserves and changes in the exchange rate."

presents the basic model and uses it to study the effect of changes in domestic monetary policy on the evolution of these three endogenous variables. The empirical implementation of the model, presented in Section II, uses quarterly data from Brazil covering the crawling-peg period, 1968–77. The parameters of the model are estimated simultaneously using a full-information maximum-likelihood method. Concluding remarks are presented in Section III.

I. The Model

Monetary Equilibrium

The model is a variant of the monetary approach to the balance of payments. Its main feature is that it accounts for different degrees of exchange-rate flexibility (from a freely floating parity to a completely fixed exchange rate). We take the case of a small country, defined as one whose international price of traded goods is exogenously determined, and we allow for the existence of "nontraded" goods, defined as goods whose price responds, at least in the short run, to domestic monetary variables.

The basic relationships of the monetary sector are

$$M_s \equiv a(R + D), \tag{1}$$

$$M_d = Pm_d, \tag{2}$$

$$m_d = f(y, \pi^e), \tag{3}$$

where M_s is the nominal supply of domestic money; a is the money multiplier; R is the foreign-exchange reserves held by the central bank; D is the domestic-credit component of the monetary base; M_d is the demand for nominal cash balances; P stands for a price index that includes traded and nontraded goods; and m_d is the real demand for money, assumed to be a function of real income,[4] y, and of the alternative cost of holding money, proxied by π^e, the expected rate of inflation.[5]

Although it is possible to postulate a mechanism of lagged adjustment, as is done in the empirical section, here we assume for

[4] We assume that the rate of growth of real income is exogenous with respect to monetary variables (we therefore assume a version of the natural rate hypothesis). Although restrictive, this assumption allows us to focus on the specific effects of money in an open economy in the context of the monetary approach. At the empirical level we found evidence supporting our assumption for the case under consideration. A complete report of the relevant causality tests is available from us on request.

[5] As it happens, the results below are valid under a large variety of assumptions about the mechanism for the formation of π^e, provided that the latter depends on lagged variables. In the empirical applications, we assume that π^e is formed by a version of rational expectations; see Sec. II.

simplicity that the money market clears in each period so that the nominal stock of money is equalized, ex post, with the demand for nominal balances. The assumption requires the existence of the flow equilibrium

$$M_s^* = M_d^*, \tag{4}$$

where the asterisk indicates the percentage rate of change of the variable. Differentiating equations (1) and (2) logarithmically, the flow-equilibrium condition can be rewritten as

$$a^* + (1 - \gamma)R^* + \gamma D^* = P^* + m_d^*, \tag{5}$$

where γ is a factor of proportionality equal to $D/(R + D)$ and P^* is the domestic rate of inflation.

The Domestic Rate of Inflation and the Balance of Payments

When traded and nontraded goods are both present, the domestic rate of inflation can be measured as a weighted average of the rate of change of the price of both kinds of goods,

$$P^* = \lambda P_T^* + (1 - \lambda)P_{NT}^*, \tag{6}$$

where P_T, P_{NT} is the price in domestic currency of traded, nontraded goods, respectively, and λ is the share of traded goods in total expenditure. In a small economy P_T^* is determined by the world rate of inflation (P_w^*) and by the variations in the effective exchange rate (ρ^*):

$$P_T^* = P_w^* + \rho^*. \tag{7}$$

The price of nontraded goods, however, may be affected by domestic factors, at least in the short run. Since an ex ante excess supply of money implies excess demand in the goods market, and if excess demand for nontraded goods varies monotonically with excess demand throughout the economy, we can expect the price of nontraded goods relative to that of traded goods to be a function of imbalance in the money market,

$$P_{NT}^* = P_T^* + \theta\Omega, \tag{8}$$

where Ω is a measure of monetary imbalance and θ is the elasticity of the relative price with respect to the monetary variable.[6]

An important feature of equation (8) is that it includes only con-

[6] The elasticity θ is a function of the elasticity of substitution between traded and nontraded goods in consumption and production and of the income elasticity of the nontraded goods. For a detailed description of the dynamics of domestic-price determination in a monetary model with traded and nontraded goods, see, e.g., Parkin (1974) and Blejer (1977).

temporaneous variables. In practice, however, it may well be that P^*_{NT} depends on lagged P^*_T and Ω reflecting persistence effects. To allow for these effects, it is possible to postulate a relationship of the form

$$(P^*_{NT} - P^*_T)_t = \theta\Omega_t + \alpha(P^*_{NT} - P^*_T)_{t-1}, \tag{9}$$

where t is a time subscript and α is the adjustment coefficient. Empirically, however, we found that our results were quite insensitive to the particular specification adopted. We therefore considered equation (8) as the relevant representation for relative-price determination. The structure of the model resulting from equation (9) and the results obtained using that specification are presented in the Appendix.[7]

As regards the measure Ω, it is important to remember that a central conclusion of the monetary approach to the balance of payments is that in a small open economy the nominal supply of money may be beyond the control of the monetary authority. Under fixed exchange rates the monetary authority can only determine the ex ante quantity of money by changing the domestic-credit component of the base or by manipulating the money multiplier. In conjunction with the flow demand for real balances generated by adjustments in the desired stock, such measures create an ex ante excess flow supply of money to which the public reacts by changing the level of the international reserve component of the base through the balance of payments and by affecting the rate of domestic inflation. The ex post nominal quantity of money in an open economy is then influenced by the public's response to ex ante conditions in the money market.[8]

It appears, therefore, that the relevant measure to account for the monetary effects on the goods market in an open economy should be an ex ante measure which does not include the endogenous reaction of the foreign component of the base. For that reason we define Ω in equation (8) as the gap (in percentage terms) between the ex ante change in the money supply (i.e., a change in the domestic-credit component of the base and in the money multiplier)[9] and changes in demand. Equation (8) can therefore be rewritten as

$$P^*_{NT} = P^*_T + \theta(\gamma D^* + a^* - M^*_d). \tag{10}$$

[7] Eq. (8) ignores possible real sector effects on the relative price of nontraded goods (e.g., real wages, income policies). Although the equation can be modified to take these factors into account, this would result in a framework far beyond our intended extension of the monetary approach to the case of a crawling peg.

[8] For time-series analysis of European data and econometric tests of the interaction between changes in domestic credit and in international reserves implied by the monetary approach, see Blejer (1979) and Leiderman (1980).

[9] Empirical examination of the assumption of exogeneity of $(\gamma D^* + a^*)$, on the basis of causality tests, indicates nonrejection of the assumption; complete test results are available from us on request.

Substituting (7) into (10) and then (7) and (10) into (6), we obtain, after some manipulation, the following expression for the rate of domestic inflation:[10]

$$P^* = \epsilon(P_w^* + \rho^*) + (1 - \epsilon)(\gamma D^* - m_d^*), \tag{11}$$

where $\epsilon = [1 + \theta(1 - \lambda)]^{-1}$.

In addition to changes in the price level, there are also changes in international reserves operating to restore monetary equilibrium. An expression for the money account of the balance of payments, which is equal to the change in the international reserves held by the central bank, may therefore be obtained by substituting P^* in equation (5)—the flow equilibrium condition for the money market—for its value in (11) and rearranging the terms:

$$(1 - \gamma)R^* = \epsilon(P_w^* + \rho^* + m_d^* - \gamma D^*). \tag{12}$$

Equations (11)–(12) present the domestic rate of inflation and the balance of payments as functions of world inflation, exchange-rate policy, and the rate of ex ante excess flow supply of money. When nontraded goods are absent ($\lambda = 1$) or when their price is not sensitive to monetary imbalance ($\theta = 0$), then $\epsilon = 1$, and the model is similar to the classical long-run formulation of the monetary approach (see Johnson 1972). In such a case, domestic monetary variables do not affect the domestic rate of inflation which, if the exchange rate is not altered, is fully determined by the world rate, and every ex ante monetary shock will lead to reserve depletion due to a balance-of-payments deficit.

The Endogeneity of the Exchange Rate in a Crawling-Peg System

Except in a fully flexible exchange-rate system (or in a managed float), the exchange rate is regarded by governments as a policy instrument, and its fluctuations are generally influenced by policy decisions aimed at one or more goals. Unlike under an adjustable-peg regime, under a crawling-peg system the exchange rate is changed frequently according to some set of rules adopted in order to attain the government's objectives.[11] The variation of the exchange rate can therefore be considered as following a sort of reaction function which reflects policy goals as well as the parameters of the model adopted and the

[10] Assuming, for simplicity, a constant money multiplier, i.e., $a^* = 0$. In the empirical section, however, changes in the money multiplier are explicitly considered.

[11] A number of proposals for the operation of the crawling peg as well as analyses of the stability of the system have recently been presented in the literature: see, e.g., Williamson (1966), Cooper (1970), Kenen (1975), Levin (1975, 1977), and Mathieson (1976).

values of the exogenous and endogenous variables considered relevant for the desired goals. The goals, and therefore the crawling rules, may vary from country to country.[12]

With a view to the empirical implementation of the model for Brazil, we postulate here that the policy objective is to avoid long-run changes in the real exchange rate and that the nominal rate (ρ) is therefore altered to maintain purchasing power parity.[13] We assume, in addition, that the full adjustment of the exchange rate may take more than one period, and we shall analyze the effects of differences in the speed of exchange-rate adjustment. The reaction function implied by the policy rule is

$$\rho_t^* = \beta \sum_{i=0}^{n} (1 - \beta)^i L^i (P^* - P_w^*)_t, \tag{13}$$

where t is a time subscript, β indicates the portion of the current differential rate of inflation transmitted to the exchange rate in the current period, and L is the lag operator ($L^i x_t = x_{t-i}$).[14]

This formulation specifically assumes that, in addition to the current-period adjustment, the exchange rate will continue to change in each subsequent period by a fraction β of the still unadjusted differential until the whole differential has been transmitted to the exchange rate. Obviously, the greater is β, the faster will the rate adjust. If $\beta = 1$ our model does not differ conceptually from the monetary-approach model of free-floating exchange rates, since the rate of depreciation is then fully determined by the domestic-foreign inflation differential (see Frenkel 1976).

The Functioning of the Model

We proceed now to solve the model for the three endogenous variables in which we are interested: the rate of inflation, the rate of

[12] Kenen (1975) analyzes in detail the relative efficiency of a number of alternative sliding-parity rules. Mathieson (1976) studies the consequences of using a welfare—instead of a balance-of-payments—objective as the guideline for the crawl

[13] The appropriateness of this assumption for Brazil is discussed in Sec. II. In a previous version of this paper a number of alternative rules were incorporated into the model, among them maintaining a given level of nominal reserves ($R^* = 0$) and maintaining a given level of real reserves ($R^* - P_w^* = 0$). Although the dynamics of the system change with the policy rule, its basic structure is not affected. For presentational convenience the functioning of the model is here confined to a single rule.

[14] A reaction function of this type follows Dean (1974) in the sense that endogenous target variables are a function of the current values of other endogenous variables; this is justified on the grounds that the latter are a plausible representation of the structure of the model and may therefore provide prior knowledge of the structural-form coefficients. This differs from the approach of earlier works such as Friedlaender (1973) where endogenous variables depend only on exogenous or lagged-endogenous variables.

change of foreign-exchange reserves, and the rate of change of the exchange rate. Combine (11) and (13) to solve for ρ_t^*,

$$\rho_t^* = \frac{\beta(1-\epsilon)}{1-\beta\epsilon-(1-\beta)L}(\gamma D^* - m_d^* - P_w^*)_t, \tag{14}$$

and substitute this in (11) and (12) to obtain

$$P_t^* = \frac{\epsilon(1-\beta)(1-L)}{1-\beta\epsilon-(1-\beta)L}(P_w^*)_t + \frac{(1-\epsilon)[1-(1-\beta)L]}{1-\beta\epsilon-(1-\beta)L}(\gamma D^* - m_d^*)_t, \tag{15}$$

and

$$(1-\gamma)R_t^* = \frac{\epsilon(1-\beta)(1-L)}{1-\beta\epsilon-(1-\beta)L}(P_w^* + m_d^* - \gamma D^*)_t. \tag{16}$$

The three endogenous variables are functions of current and lagged values of foreign inflation and domestic monetary variables. We can consider now which type of monetary policy will equalize the domestic rate of inflation to the world rate. Because the coefficients of P_w^* and $(\gamma D^* - m_d^*)$ in equation (15) add to unity, P^* will equal P_w^* when the monetary authority expands the money supply at the rate necessary to satisfy the growth in real demand and to replace the depreciated value of the nominal stock. This is achieved when the exogenous component of the supply of money (domestic credit) expands at a rate that exceeds the growth in the demand for real balances by the world rate of inflation, that is, when the ex ante excess flow supply of money is equal to the world inflation rate,

$$(\gamma D^* - m_d^*) = P_w^*, \tag{17}$$

which implies

$$P^* = P_w^* \tag{18}$$

and

$$\rho^* = R^* = 0. \tag{19}$$

To illustrate the functioning of the model further we can analyze the effects of a domestic-credit shock. To do so, it is convenient to consider an economy whose initial equilibrium satisfies equations (18)–(19). Starting from such an equilibrium position, and as long as β and ϵ are smaller than unity, an acceleration in the rate of expansion of the domestic-credit component of the monetary base, $\Delta(\gamma D^*)$, will cause the rate of domestic inflation to depart from the world rate (here assumed constant), the exchange rate will rise, and a balance-of-payments deficit will be created.

If the rate of growth of domestic credit is sustained at the new, higher level, the exchange rate will eventually adjust to account fully for the differential in the inflation rates, which will then be equal to $\Delta(\gamma D^*)$.[15] Once the changes in the exchange rate fully reflect the differential rate of inflation, the excess supply of money created by the government in each period is fully eliminated through an increase in domestic prices, and no further flows of foreign exchange reserves occur. This process is illustrated in figure 1 for different values of β and ϵ.[16] As β increases, for given values of ϵ, the exchange rate adjusts faster to the monetary shock, and domestic inflation both diverges faster from world inflation and converges faster to its new equilibrium rate. As can be observed by comparing paths III and II in figure 1, the balance of payments deteriorates less and returns to equilibrium more rapidly (implying a smaller total loss of reserves) as β increases. When $\beta \rightarrow 1$, the system approaches a flexible-exchange-rate model, like the one presented by Frenkel (1976), in which domestic inflation is always independent of world inflation and the balance of payments is always zero. Letting $\beta = 1$ in equations (14)–(16), we obtain

$$P_t^* = (\gamma D^* - m_d^*)_t, \tag{20}$$

$$\rho_t^* = (\gamma D^* - m_d^* - P_n^*)_t = (P^* - P_w^*)_t, \tag{21}$$

$$R_t^* = 0. \tag{22}$$

The speed of adjustment will also increase, for given values of β, the lower the ϵ, that is, the higher the share of nontraded goods in expenditures, λ, or the higher the elasticity of relative prices with respect to monetary imbalance, θ. The more rapid adjustment of the rate of exchange when ϵ falls (which also implies a faster acceleration of the rate of inflation although a smaller loss of reserves during the adjustment process) is illustrated by the comparison of paths II and I in figure 1.

Thus the results indicate that under a crawling-peg system a small country can choose its own long-run rate of inflation independently from the rest of the world. However, balance-of-payments deficits and surpluses are experienced in the process of adjustment, owing to purchasing power disparities which arise because the exchange rate takes some time to adjust fully.

We turn now to the analysis of the empirical results obtained by applying the model described above to the case of Brazil.

[15] That is, the difference between the new rate of ex ante excess supply of money $[\gamma D^* + \Delta(\gamma D^*) - m_d^*]$ and the world rate of inflation.

[16] For presentational simplicity the figure does not consider the effects of the temporary reduction in m_d^* as inflationary expectations accelerate. The pattern of adjustment is, however, very similar when these effects are taken into account.

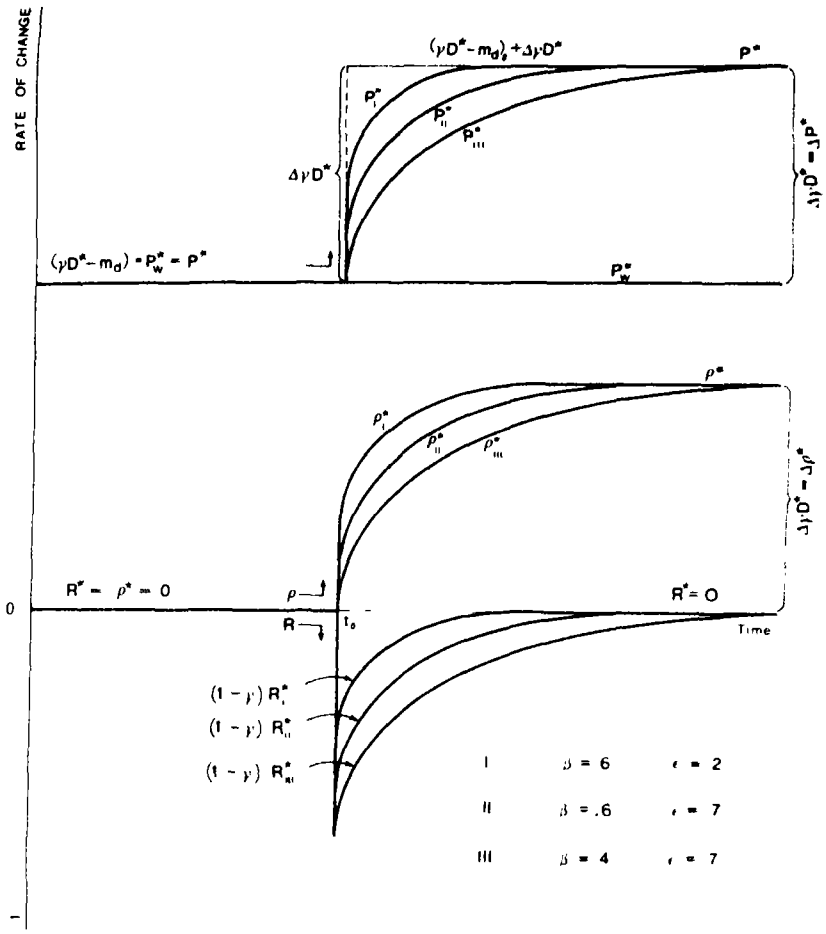


FIG. 1

II. Empirical Investigation

Estimation Procedures

In this section, the model developed above is restated in econometric form. For estimation purposes, we work with the following system:

$$\log m_t = \delta_1 \log y_t - \delta_2 \pi^e_t + \delta_3 \log m_{t-1} + \mu_{1t}, \tag{3'}$$

$$P^*_t = \epsilon(P^*_w + \rho^*)_t + (1 - \epsilon)(a^* + \gamma D^* - m^*_d)_t + \mu_{2t}, \tag{11'}$$

$$(1 - \gamma)R^*_t = \epsilon(P^*_w + \rho^* + m^*_d - \gamma D^* - a^*)_t + \mu_{3t}, \tag{12'}$$

$$\rho^*_t = \beta(P^* - P^*_w)_t + \beta(1 - \beta)(P^* - P^*_w)_{t-1} + \beta(1 - \beta)^2(P^* - P^*_w)_{t-2} + \mu_{4t}. \tag{13'}$$

Equation (3′) is the empirical counterpart of the model's demand for real money balances. The specification is semilogarithmic, in that the log of real money balances depends on the log of real income and expected inflation. As the model is applied to quarterly data, we postulate a process of gradual adjustment of real balances to their optimum value, and that is the rationale for the appearance of lagged money balances in the equation.[17] Equations (11′)–(13′) are stochastic variants of equations (11)–(13) above.[18]

The following stochastic specification is adopted: It is assumed that μ_i has a first-order autoregressive representation $\mu_{it} = \phi_i\mu_{it-1} + v_{it}$ (for $i = 1, 2, 3, 4$), where ϕ_i is the autoregressive parameter, and v_{it} is the error term. Furthermore, we assume that the error vectors, $v_{1t}, v_{2t}, v_{3t}, v_{4t}$ (for $t = 1, \ldots, T$) have zero mean and a constant variance-covariance matrix Γ; also, they are serially uncorrelated and drawn from a multivariate normal distribution.

Equations (3′) and (11′)–(13′) constitute a simultaneous system, which is linear in the variables but nonlinear in (some of) the parameters. Given these characteristics, consistent estimation requires the use of some simultaneous-equation method. We have used a full-information maximum-likelihood (FIML) estimator; FIML utilizes all the a priori restrictions on the system to estimate the coefficients of interest (in our case, $\epsilon, \beta, \phi_i, \delta_i$) by maximizing the model's likelihood function.[19]

To make the estimation operational, a time series of m^*_{dt}, the rate of change of real money demand, is required. As money demand is simultaneously estimated with the other equations of the model, the complete system is estimated under the restriction that m^*_{dt} corresponds to the first difference of the fitted values of equation (3′). In addition, note that the estimation of money demand requires a proxy for π^f. This proxy is constructed by assuming a version of the rational expectations hypothesis. Specifically, it is postulated that agents form their inflation expectations on the basis of the least-squares prediction of inflation, conditional on the set of available information. We have

[17] Formally, the demand for money is given by $\log m_{d_t} = \delta_1' \log y_t - \delta_2'\pi_t^f$. Money demand equations of this form have previously been estimated for Brazil by Silveira (1973) and Khan (1977). We assume an adjustment mechanism of the form $(\log m_t - \log m_{t-1}) = h(\log m_{dt} - \log m_{t-1})$, $0 < h < 1$. See Griliches (1967) for a discussion of the conditions under which adjustment mechanisms of this class are likely to be optimal.

[18] As can be seen, we are now taking into account changes in the money multiplier. Note also that we consider only three terms in the exchange-rate equation (this issue is discussed further below).

[19] See Hendry (1971). The parameter estimates reported below were obtained by using Wymer's (1978) computer program, RESIMUL. The program uses a Newton-Raphson iterative procedure, beginning with arbitrarily given initial values of the parameters.

assumed that the relevant information set includes lagged values of the variables which, according to the model, determine the rate of inflation.[20]

4 Crawling-Peg Experience: The Case of Brazil

Brazil's adoption of the crawling-peg regime can be considered as an extension of indexation to the external sector. Starting in 1964, Brazil engaged in one of the most comprehensive indexation programs anywhere. The financial sector was the first to be indexed. Exchange-rate indexation was introduced in August 1968.[21]

During the postwar period, Brazil had experienced a variety of exchange-rate regimes. Until 1953, exchange controls and licensing were widely used. In 1953 an exchange auction system with multiple exchange rates was introduced. A process of unifying the rates was initiated in 1961, and since then the country has had a pegged exchange rate. Since 1968, when the crawling peg was adopted, new selling and buying rates for the cruzeiro in relation to the U.S. dollar are announced by the central bank at frequent intervals (on average, the rate was changed over 10 times a year).

Despite the fact that there are no explicitly stated rules for the crawl, it is evident that the method in effect applies a purchasing power parity clause to the domestic price of foreign exchange, and the evolution of both domestic and foreign price levels is taken into account in adjusting the exchange rate. In this way, as emphasized by the Brazilian authorities, it is possible to neutralize the harmful effects of domestic inflation on the competitiveness of Brazilian products in international markets and, hence, on Brazil's balance of payments.[22]

Although inflation differentials are certainly an important factor in determining the crawl, it was not the only one. However, it is evident from the data (see Lemgruber 1977, tables 1 and 6; and Moura da Silva 1977, table 2) that since 1968 the variations in the exchange rate have followed purchasing power parity rates closely, suggesting that exchange-rate changes have been mainly determined by the differ-

[20] Specifically, π^e was constructed from the fitted values of a regression of the rate of inflation on lagged values of itself, foreign inflation, domestic-credit growth, exchange-rate devaluations, and real income growth. For each of these variables two lagged values were included in the equation.

[21] On the various aspects of indexation in Brazil, see Nadiri and Pastore (1977). A comprehensive study of the recent Brazilian economic experience is presented by Lemgruber (1977). On the functioning of the crawling peg and its effects on the economy, see Moura da Silva (1977). See also Lara-Resende (1978).

[22] This appears to be the main justification put forward by the authorities for the adoption of the crawling peg; on this issue see Lemgruber (1977) and Moura da Silva (1977)

ence between domestic and foreign inflation.[23] The specification of the exchange-rate reaction function above (eq. [13]) is intended precisely to capture this feedback process. Such a specification imposes restrictions on the data; econometric tests of these restrictions as well as the results from estimating the model are reported below.

Empirical Results

We now report the estimates of the parameters of equations (3′) and (11′)–(13′), as well as of the derived reduced-form equations. All estimates are based on quarterly data for Brazil. The sample period, III/1968 through IV/1977, contains 38 observations. Most of the variables are represented by readily available series.[24] The results are presented in table 1. All the parameter estimates have the proper sign and are reasonably well determined; in most cases asymptotic t ratios are well above 2.[25]

All the coefficients in the estimated money-demand equation (3′) have the expected signs and are significantly different from zero. The estimated coefficients are fairly similar to those reported for Brazil by Silveira (1973) and Khan (1977). Our parameter estimates imply a long-run income elasticity of real money demand of 1.307 and a long-run elasticity or real money demand with respect to expected inflation of −0.77.

The results for (11′)–(13′) indicate that both foreign inflation and excess supply in the money market significantly affect the determination of the domestic rate of inflation. These two factors, world inflation and monetary imbalance, also significantly affect the balance of payments, as shown by equation (12′). Specifically, an acceleration in the rate of domestic-credit creation over and above the rate of growth in money demand will increase domestic inflation and reduce international reserves. The overall effect of such an increase, however, cannot be established solely from equations (11′) and (12′). This

[23] The exchange rule has been followed with a considerable degree of consistency. Almonacid and Pastore (1977) present some evidence to that effect and criticize the lack of flexibility—the rule has not been altered to deal with the oil crisis and with the mounting external debt of Brazil. Similar evidence is also presented by Connolly and da Silveira (1979, pt. 3), who found that purchasing power parity holds very well, particularly after 1962.

[24] The following series were used for P, R, and M: consumer price index, international reserves of the central bank, and the money supply; D was generated by subtracting R from the monetary base; ρ is the market exchange rate, selling rate (average over the quarter) (all the above were taken from *International Financial Statistics*, various issues); y is represented by the quarterly real income series constructed by Wachter (1976). (The series were extended through 1977 using Wachter's method.)

[25] Asymptotic t ratios are defined as coefficient values divided by asymptotic standard errors.

TABLE 1

FULL-INFORMATION MAXIMUM-LIKELIHOOD ESTIMATES OF THE MODEL

(3′) $$\log m_t = \underset{(0.049)}{0.162} \log y_t - \underset{(0.229)}{1.700}\, \pi_t^e + \underset{(0.045)}{0.876} \log m_{t-1},$$

$$\text{MSE} = 0.0018,\ \phi_1 = \underset{(0.042)}{-0.613},\ C_1 = \underset{(0.297)}{0.873};$$

(11′) $$P_t^* = \underset{(0.018)}{0.958}(P_w^* + \rho^*)_t + \underset{(0.018)}{0.042}(a^* + \gamma D^* - m_d^*)_t,$$

$$\text{MSE} = 0.0003,\ \phi_2 = \underset{(0.110)}{-0.082},\ C_2 = \underset{(0.0012)}{0.002};$$

(12′) $$(1 - \gamma)R_t^* = \underset{(0.018)}{0.958}(P_w^* + \rho^* + m_d^* - \gamma D^* - a^*)_t,$$

$$\text{MSE} = 0.0021,\ \phi_3 = \underset{(0.045)}{0.251},\ C_3 = \underset{(0.039)}{-0.038};$$

(13′) $$\rho_t^* = \underset{(0.106)}{0.526}(P^* - P_w^*)_t + \underset{(0.005)}{0.249}(P^* - P_w^*)_{t-1} + \underset{(0.029)}{0.118}(P^* - P_w^*)_{t-2},$$

$$\text{MSE} = 0.0005,\ \phi_4 = \underset{(0.106)}{-0.050},\ C_4 = \underset{(0.003)}{-0.008}.$$

$$\Gamma = \begin{bmatrix} 0.0018 & & & \\ -0.0004 & 0.0006 & & \\ 0.0009 & 0.0004 & 0.0029 & \\ 0.0002 & -0.0005 & -0.0005 & 0.0005 \end{bmatrix}$$

NOTE.—See text for notation. Numerals in parentheses are asymptotic standard errors of regression coefficients. Γ is the estimated variance-covariance matrix of residuals, C_i ($i = 1, 2, 3, 4$) is the estimated constant term, ϕ_i is the autoregression coefficient, and MSE is the mean square error.

is so because the divergence of domestic from world inflation will, as indicated by equation (13′), depreciate the exchange, which in turn will further increase the rate of inflation and reduce the rate at which foreign reserves are depleted. In order to analyze the full effect of changes in the rate of growth of domestic credit (as well as in the other exogenous variables), reduced-form coefficients should be considered. These coefficients are reported below.

The parameter estimates in equation (13′)—the exchange-rate reaction function—indicate that more than 50 percent of the foreign-domestic inflation differential is transmitted to the exchange rate within the current quarter, and an additional 25 percent in the next. After only three quarters the full differential is completely reflected by the exchange-rate adjustment. In this context, it is important to note that equation (13) embodies a restriction regarding the effects of current and lagged values of the domestic-foreign inflation differential on the exchange rate. Specifically, a pattern of geometrically decaying weights was postulated. The results reported in table 1 are derived under the imposition of this restriction. To test

Appendix

In this Appendix we derive the structural and reduced-form equations of the model when persistence effects are allowed in the price equation for nontraded goods. The structural equations using this specification are then estimated and the results are reported below.

Replacing equation (8) by equation (9), we can express the rate of change of nontraded goods prices as

$$(P^*_{NT})_t = (P^*_T)_t + \frac{\theta}{1 - \alpha L}\Omega_t. \tag{A1}$$

Adopting now the definition of Ω_t given in Section I, we obtain equation (A2), the counterpart of equation (10),

$$(P^*_{NT})_t = (P^*_T)_t + \frac{\theta}{1 - \alpha L}(\gamma D^* + a^* - M^*_d)_t. \tag{A2}$$

Using equations (7) and (A2), the rate of domestic inflation, equation (6), can be written as

$$P^*_t = \epsilon'(P^*_u + \rho)_t + (1 - \epsilon')(\gamma D^* + a^* - m^*_d)_t, \tag{A3}$$

where ϵ' is now a polynomial in the lag operator,

$$\epsilon' = \frac{1 - \alpha L}{1 - \alpha L + \theta(1 - \lambda)}. \tag{A4}$$

Postulating the same crawling rule as used in the text, that is, equation (13), we can obtain the reduced forms of the model,[27]

$$\rho^*_t = \frac{\beta(1 - \epsilon')}{1 - \beta\epsilon' - (1 - \beta)L}(\gamma D^* + a^* - m^*_d - P^*_u)_t, \tag{A5}$$

$$P^*_t = \frac{\epsilon'(1 - \beta)(1 - L)}{1 - \beta\epsilon' - (1 - \beta)L}(P^*_u)_t + \frac{(1 - \epsilon')[1 - (1 - \beta)L]}{1 - \beta\epsilon' - (1 - \beta)L}(\gamma D^* + a^* - m^*_d)_t, \tag{A6}$$

$$(1 - \gamma)h^*_t = \frac{\epsilon'(1 - \beta)(1 - L)}{1 - \beta\epsilon' - (1 - \beta)L}(P^*_u + m^*_d - \gamma D^*)_t. \tag{A7}$$

The econometric forms of the model, in conjunction with the demand for money, equation (3′), are estimated by the procedure described in Section II. The parameter estimates do not differ much from the model in which this type of persistence is ignored.

For comparison and completeness we now report the estimates obtained with the extended model:

[27] The reduced forms could be expressed in terms of the underlying parameters, λ, α, and θ by substituting the RHS of (A4) for ϵ'. This results in rather cumbersome expressions, as, e.g., in eq. (A5)

$$\rho^*_t = \frac{\beta\theta(1 - \lambda)}{(1 - \alpha L)(1 - \beta)(1 - L) + \theta(1 - \lambda)(1 - \beta)L}(\gamma D^* + a^* - m^*_d - P^*_u)_t.$$

$$\log m_t = \underset{(0.063)}{0.224} \log y_t - \underset{(0.300)}{2.224}\pi_t^r + \underset{(0.058)}{0.828} \log m_{t-1}, \tag{A3'}$$

$$\rho_t^* = \underset{(0.099)}{0.618}(P^* - P_w^*)_t + \underset{(0.023)}{0.236}(P^* - P_w^*)_{t-1} + \underset{(0.032)}{0.090}(P^* - P_w^*)_{t-2}, \tag{A8}$$

$$P_t^* = \underset{(0.018)}{0.951}(P_w^* + \rho^*)_t + \underset{(0.018)}{0.049}(a^* + \gamma D^* - m_d^*)_t, \tag{A9}$$

$$(1 - \gamma)R_t^* = \underset{(0.018)}{0.951}(P_w^* + \rho^* + m_d^* - \gamma D^* - a^*)_t. \tag{A10}$$

References

Almonacid, Rubén D., and Pastore, Alfonso C. "El tipo de cambio, la crisis del petroleo y la deuda externa de Brasil." *Cuadernos econ* (Santiago, Chile) 14, no. 43 (December 1977): 109–27.

Artus, Jacques R. "Exchange Rate Stability and Managed Floating: The Experience of the Federal Republic of Germany." *International Monetary Fund Staff Papers* 23 (July 1976): 312–33.

Blejer, Mario I. "The Short-Run Dynamics of Prices and the Balance of Payments." *A.E.R* 67 (June 1977): 419–28.

———. "On Causality and the Monetary Approach to the Balance of Payments: The European Experience." *European Econ Rev.* 12 (July 1979): 289–96.

Boyer, Russell S. "Optimal Foreign Exchange Market Intervention." *J.P.E* 86, no. 6 (December 1978): 1045–55

Connolly, Michael, and da Silveira, José Dantas. "Exchange Market Pressure in Postwar Brazil: An Application of the Girton-Roper Monetary Model." *A.E.R.* 69 (June 1979). 448–54

Cooper, Richard N. "Sliding Parities: A Proposal for Presumptive Rules." In *Approaches to Greater Flexibility of Exchange Rates: The Burgenstock Papers*, edited by George N. Halm. Princeton, N.J.: Princeton Univ. Press, 1970.

Dean, James W. "Problems in the Specification and Interpretation of Central Bank Reaction Functions." *Econ. and Soc. Rev.* 5 (July 1974): 431–43.

Dornbusch, Rudiger. "Devaluation, Money, and Nontraded Goods." *A.E.R.* 63 (December 1973): 871–80.

———. "The Theory of Flexible Exchange Rate Regimes and Macroeconomic Policy." *Scandinavian J. Econ.* 78 (May 1976): 255–75.

Frenkel, Jacob A. "A Monetary Approach to the Exchange Rate: Doctrinal Aspects and Empirical Evidence." *Scandinavian J. Econ* 78 (May 1976): 200–24.

———. "The Demand for International Reserves under Pegged and Flexible Exchange Rate Regimes and Aspects of the Economics of Managed Float." In *The Functioning of Floating Exchange Rates: Theory, Evidence, and Policy Implications*, edited by David Bigman and Teizo Taya. Cambridge, Mass.: Ballinger, 1980.

Frenkel, Jacob A., and Clements, Kenneth W. "Exchange Rates in the 1920's: A Monetary Approach." In *Development in an Inflationary World*, edited by M. J. Flanders and Assaf Razin. New York: Academic Press, in press.

Frenkel, Jacob A., and Johnson, Harry G., eds. *The Monetary Approach to the*

Balance of Payments. London: Allen & Unwin; Toronto: Univ. Toronto Press, 1976.

———, eds. *The Economics of Exchange Rates: Selected Studies*. Reading, Mass.: Addison-Wesley, 1978.

Friedlaender, Ann F. "Macro Policy Goals in the Postwar Period: A Study in Revealed Preference." *Q.J.E.* 87 (February 1973): 25–43.

Girton, Lance, and Roper, Don. "A Monetary Model of Exchange Market Pressure Applied to the Postwar Canadian Experience." *A.E.R.* 67 (September 1977): 537–48.

Griliches, Zvi. "Distributed Lags: A Survey." *Econometrica* 35 (January 1967): 16–49.

Hahn, Frank H. "The Monetary Approach to the Balance of Payments." *J. Internat. Econ.* 7 (August 1977): 231–49.

Hendry, David F. "Maximum Likelihood Estimation of Systems of Simultaneous Regression Equations with Errors Generated by a Vector Autoregressive Process." *Internat. Econ. Rev.* 12 (June 1971): 257–72.

International Monetary Fund (IMF). *The Monetary Approach to the Balance of Payments*. Washington: IMF, 1977.

Johnson, Harry G. "The Monetary Approach to the Balance-of-Payments Theory." *Further Essays in Monetary Theory*. London: Allen & Unwin, 1972.

——— "The Monetary Approach to the Balance of Payments: A Nontechnical Guide." *J. Internat. Econ.* 7 (August 1977): 251–68.

Kenen, Peter B. "Floats, Glides and Indicators: A Comparison of Methods for Changing Exchange Rates." *J. Internat. Econ.* 5 (May 1975): 107–51.

Khan, Mohsin S. "Variable Expectations and the Demand for Money in High-Inflation Countries." *Manchester School Econ. and Soc. Studies* 45 (September 1977): 270–93.

Kouri, Pentti J. K. "Exchange Rate Expectations, and the Short-Run and the Long-Run Effects of Fiscal and Monetary Policies under Flexible Exchange Rates." Paper presented at the conference on Monetary Mechanisms in Open Economics, Helsinki, August 1975.

Lara-Resende, Andre. "Interest Parity and the Brazilian Crawling Peg: The Official and Parallel Markets." Mimeographed. Cambridge, Mass.: Massachusetts Inst. Technol., 1978.

Leiderman, Leonardo. "Relationships between Macroeconomic Time-Series in a Fixed-Exchange-Rate Economy." *European Econ. Rev.* 14 (July 1980): 61–77.

Lemgruber, Antonio C. "Inflation in Brazil." In *Worldwide Inflation: Theory and Recent Evidence*, edited by Lawrence B. Krause and Walter S. Salant. Washington: Brookings Inst., 1977.

Levin, Jay H. "Monetary Policy and the Crawling Peg." *Econ. J.* 85 (March 1975): 20–32.

———. "Reserve Stocks as External Targets and the Stability of Alternative Exchange Rate Systems." *Rev. Econ. Studies* 44 (February 1977): 59–70.

Magee, Stephen P. "The Empirical Evidence on the Monetary Approach to the Balance of Payments and Exchange Rates." *A.E.R. Papers and Proc.* 66 (May 1976): 163–70.

Mathieson, Donald J. "Is There an Optimal Crawl?" *J. Internat. Econ.* 6 (May 1976): 183–202.

Moura da Silva, Adroaldo. "The Basis of the Minidevaluation Policy." *Explorations Econ. Res.* 4 (Winter 1977): 101–16.

Mundell, Robert A. *International Economics*. New York: Macmillan, 1968.

———. *Monetary Theory: Inflation, Interest, and Growth in the World Economy.* Pacific Palisades, Calif.: Goodyear, 1971.

Mussa, Michael. "The Exchange Rate, the Balance of Payments, and Monetary and Fiscal Policy under a Regime of Controlled Floating." *Scandinavian J. Econ.* 78 (May 1976): 229–48.

Nadiri, M. Ishaq, and Pastore, Alfonso C., eds. *Indexation: The Brazilian Experience.* Special issue of *Explorations Econ. Res.*, vol. 4 (Winter 1977).

Parkin, Michael. "Inflation, the Balance of Payments, Domestic Credit Expansion, and Exchange Rate Adjustments." In *National Monetary Policies and the International Financial System*, edited by Robert Z. Aliber. Chicago: Univ. Chicago Press, 1974.

Silveira, Antonio M. "The Demand for Money: The Evidence from the Brazilian Economy." *J. Money, Credit and Banking* 5 (February 1973): 113–40.

Wachter, Susan M. *Latin American Inflation: The Structuralist-Monetarist Debate.* Lexington, Mass.: Heath, 1976.

Whitman, Marina V. N. "Global Monetarism and the Monetary Approach to the Balance of Payments." *Brookings Papers Econ. Activity*, no. 3 (1975), pp. 491–536.

Williamson, John H. *The Crawling Peg.* Princeton Essays in International Finance no. 50. Princeton, N.J.: Princeton Univ. Press, 1966.

Wymer, C. *Computer Programs: Resimul Manual.* Washington: IMF, 1978

Does Federalism Matter? Political Choice in a Federal Republic

Susan Rose-Ackerman

Yale University

This paper builds upon some well-known facts about state government to generate new conclusions about social choice on the national level of a federal republic. Citizens vote against national laws that restrict their state's ability to export costs but support laws that reduce the costs imposed on them. Individuals may seek to extend the laws passed in some states to the entire nation or may oppose preemptive laws because they benefit from variety. Since these motivations are absent in a unitary system, national support for a law will depend upon whether a unitary or a federal structure prevails.

I. Introduction

This paper builds upon some well-known facts about state government to generate new conclusions about social choice on the national level of a federal republic. The central feature of state government behavior I shall exploit is the incentive each state has to improve its own position by imposing costs on the residents of other states. This search for local advantage takes many forms. On the one hand, states may impose taxes and regulations which are borne by out-of-state residents; on the other hand, they may try to attract investment from other states by providing tax breaks and special public services.[1]

This paper was presented at the Workshop on Analytical Urban Economics at Queen's University, Kingston, June 1978. I wish to thank Bruce Ackerman, the workshop participants, and the referee of this *Journal* for helpful comments.

[1] These incentives are stressed by Walker (1969, p. 890) and Posner (1977, chap. 26). McLure (1967) estimates that about one-quarter of all state taxes are exported. Externalities such as air and water pollution are emphasized by Breton (1965), Olson (1969), Tullock (1969), Rothenberg (1970), and Oates (1972). These authors discuss how a federal system can trade off interjurisdictional variety against the internalization of externalities through boundary definition, the assignment of functions to levels of government, and intergovernmental grants.

(*Journal of Political Economy*, 1981, vol. 89, no. 1)
© 1981 by The University of Chicago. 0022-3808/81/8901-0008$01.50

Whatever their particular character, these interstate spill-ins and spill-outs can alter the substance of national legislation. Citizens will vote against laws that restrict their state's ability to export costs but will support laws that reduce the costs imposed on them by their own and other states' choices. Since these strategic motivations are absent in a unitary system, national support for particular laws will depend upon whether a unitary or a federal structure prevails.

Despite its potential importance, the impact of federalism on the strategic position of voters has been largely ignored by social scientists. In his excellent review of the literature, for example, William Riker (1975) argues that federalism has no important influence on substantive national outcomes, suggesting that it does no more than delay the passage of national legislation opposed by a substantial minority of state legislatures.[2] Although Riker's conclusion is partly an empirical proposition, it is also based on an implicit model of how government systems behave. Riker argues (1975, pp. 155–56) that the underlying distribution of tastes in the population determines political outcomes. Since federalism per se has no impact on tastes, it will not, therefore, affect national political choices in the long run. The present paper challenges these theoretical underpinnings.

Since I am concerned not with an exhaustive taxonomy but with the relationship between political power and political structure, I concentrate on simple models. The system I discuss has only two "layers"—national and state.[3] States have fixed boundaries and cover the nation's entire geographical area so that every citizen lives in only one state. I will contrast the legislative choices of a unitary government with those of a "hierarchical" federal system where higher-level governments can always preempt the legislative choices of lower-level governments. If the superior government has taken no affirmative action, however, the statutes of inferior governments are binding.[4] I

[2] Riker (1975) proposes an imaginary experiment in which matched pairs of unitary and federal systems are examined to discover if there are important public policy differences. He hypothesizes that public policy within each pair will be "remarkably similar regardless of federalism" (p. 144). Furthermore, he argues that in the United States federalism delayed national regulation of business (p. 154) and helped perpetuate racist acts (pp. 154–56) but had no long-run impact on the character of national legislation.

[3] For an attempt to justify a three-layered system, see Ylvisaker (1959). With the exception of Wechsler (1968), Posner (1977), and Winter (1977), legal commentators have all but ignored the issues discussed in this paper and have instead concentrated on the relationship between state and federal courts. For a recent article in this tradition, see Cover and Aleinikoff (1977).

[4] This model should be contrasted with others which have a strict division of functions between high- and low-level governments. For example, see Wheare (1953, pp. 32–33). Recently, concern for a strict division of authority has given way to scholarship which recognizes the importance of interactions between levels of government. See, for example, Grodzins (1960, 1966) and Elazar (1962).

assume that both state and national governments are direct democracies where the only political actors are individuals and where all decisions are made by majority vote. Political issues are separable and well defined, and they appear to voters as simple dichotomous choices between maintenance of the status quo and a change.[5] The analysis concentrates on legislative choices. I do not discuss administrative or judicial issues or the possibility that the national government will seek to administer programs through state government agencies.

Section II shows that federalism "matters" in a political system where capital can move but people cannot. Section III considers the impact of permitting interstate migration of voters, and Section IV shows that federalism can matter even in the long run when states respond to the legislative choices of other states.

II. Federalism Does Matter

The major difference between a unitary system and a hierarchical federal structure is the characterization of the status quo. Let y = the status quo in a unitary system, and let x = a proposed, exogenously defined, legislative change. Suppose, for example, that in the status quo casino gambling is illegal. The federal government, however, proposes to make such gambling legal and to levy a tax on the profits.

In a hierarchical federal system, some states have passed legislation that is similar to x. Others have a status quo that is similar to y. Let us call the former x-type states and the latter y-type states. Continuing the example, x-type states permit casino gambling and levy a tax on earnings, while casino gambling is illegal in y-type states. Assume that if all states are x-type states, then the state laws taken together would have the same impact as x. Similarly, assume that if no states have passed the x-type law, the status quo is the same as in a unitary system. Thus, in my example the state gambling laws are duplicates of x and y, respectively. This assumption is not essential to the analysis, but it simplifies the exposition considerably.

Because of interstate spillovers of costs and benefits, individuals' preferences over x and y will not necessarily be the same as their preferences for alternative state legal regimes. In the gambling example, some people may favor a statewide gambling law and vote

[5] These assumptions permit me to look at the vote on a single issue and to avoid voting cycles in which majority rule does not produce a determinant outcome (Sen 1970, p. 38). The actual relationship between state statutes and federal law in the same area is, of course, considerably more complex. In reality, the relationship is seldom obvious and federal courts are frequently called upon to sort out the overlapping authority of state and federal statutes (see Tribe 1978, pp. 378–86).

against x at the national level. A state's residents gain when out-of-staters come to gamble and consume tourist services. The state would lose this advantage if gambling were legal in every state. In a unitary system, however, these same individuals might favor x over y since they would have no special advantage to protect.

A voter's preference for x over a system with a variety of state laws will often depend upon the number and location of x-type states. The economic advantages of permitting gambling in one's state of residence are larger the smaller the number of other states which also permit gambling. The benefits are also higher the more geographically distant are the other jurisdictions with legalized gambling. Let $g_i = z_i$ if state i has passed an x-type law, $g_i = w_i$ if state i is a y-type state, and let H = total number of states. Then, the status quo in a federal system is a vector $G = (g_1, \ldots, g_i, \ldots, g_H)$ where $g_i = z_i$ or w_i. Thus we can summarize the situation of those living in x-type states as $z(G)$ and the situation of those in y-type states as $w(G)$.

In this portion of the analysis, I assume that migration across state lines is impossible, so that the only way individuals can affect the legal regime under which they live is to vote for new state or federal laws. Capital, however, is free to move between jurisdictions. Given these assumptions, we can now compare the way voters with different preferences will cast their ballots in federal and unitary republics. Turning first to a unitary system, individual preferences can be straightforwardly translated in a social choice. Law x passes if and only if:

$$N(xRy) > N(yRx), \tag{1}$$

where R is the binary relation of "weak preference" ("at least as good as"). Then $N(aRb)$ is the number of people for whom aR_jb where R_j is the weak preference relation for individual j and R_j is reflexive, transitive, and complete.[6]

In a hierarchical federal system, under the assumption that only the current level of G is relevant, x passes at the national level if and only if:

$$N_1[xRz(G)] + N_2[xRw(G)] \geq N_1[z(G)Rx] + N_2[w(G)Rx], \tag{2}$$

where $N_1(aRb)$ is the number of people living in x-type states for whom aR_jb and $N_2(aRb)$ is the number of people living in y-type states for whom aR_jb.

Expressions (1) and (2) need not yield the same results. Individuals who would prefer x to y need not also prefer x to $z(G)$ or $w(G)$. If a

[6] In Sen's terminology (1970, p. 9), R_j is an ordering. The notation in (1) follows Sen (1970, p. 71).

TABLE 1

	$xRz(G)$	$z(G)Rx$	$xRw(G)$	$w(G)Rx$
xRy	A_1 (favor x)	A_3 (positive spill-ins to x types)	B_1 (favor x)	B_3 (positive spill-ins to y types)
yRx	A_2 (negative spill-ins to x types)	A_4 (oppose x)	B_2 (negative spill-ins to y types)	B_4 (oppose x)

state is able to export the costs of a program, then its citizens are likely to oppose preemptive federal laws even though they favor x-type legislation in their own state and would favor x over y in a unitary system. Similarly, if citizens can benefit from the legislative initiatives of other states then they may vote against x although they have xR_Uy with a unitary government. The gambling example falls in the first category. Public health or pollution control laws in neighboring states may produce the second voting pattern. In fact, individual preferences may produce any of the possible rankings of x, y, and $z(G)$ and x, y, and $w(G)$; and one can thus divide the population into groups depending upon their preferences and the states in which they live. For a given G, let $A(G)$ = set of people in x-type states and let $B(G)$ = everyone else.

People living in x-type states have $xRz(G)$ or $z(G)Rx$ or both. Those in set $B(G)$ have $xRw(G)$ or $w(G)Rx$ or both. In a unitary system voters have xRy or yRx or both. Table 1 is a matrix of the possible taste combinations. The entries in the matrix are the sets of people in each taste class. Thus $A(G) = A_1(G) \cup A_2(G) \cup A_3(G) \cup A_4(G)$ and $B(G) = B_1(G) \cup B_2(G) \cup B_3(G) \cup B_4(G)$.[7] In expression (1), $N(xRy) = N(A_1 \cup A_3 \cup B_1 \cup B_3)$. In expression (2), $N_1[xRz(G)] = N_1(A_1 \cup A_2)$ and $N_2[xRw(G)] = N_2(B_1 \cup B_2)$. Thus given G, the vote for x in a unitary system, $N(xRy)$, is greater than, equal to, or less than the vote in a hierarchical system, $N_1[xRz(G)] + N_2[xRw(G)]$, as:

$$N(A_3 \cup B_3) \gtreqless N(A_2 \cup B_2). \tag{3}$$

Expression (3) shows that federalism "matters." The size of the vote in favor of the national law may be higher or lower in a hierarchical federal system than in a unitary system. A law which passes in one system may be defeated in another. The practical importance of this result, however, depends upon whether sets A_3, B_3 and A_2, B_2 represent common preference configurations. Sets A_3 and B_3 include indi-

[7] The symbol $\cup$ stands for the union of sets, i.e., the set of people belonging to either set.

viduals who prefer x to y as the national law in a unitary system but who each favor conditions in their own states over x.[8] Sets A_2 and B_2 are the reverse of A_3 and B_3. They include individuals who would oppose x in a unitary system but favor preemptive legislation because it keeps them from losing at the expense of other citizens.

If sets A_2 and B_2 are large relative to sets A_3 and B_3, then a paradoxical result is possible. Although federalism is often justified as a way to preserve diversity and prevent the centralization of power, (3) implies that the existence of low-level governments may cause some voters to favor central control of an activity that they would otherwise have wanted unregulated. Without constitutional limits, a majority of citizens may be so unhappy with the mixture of independent state choices under federalism that they favor a more powerful central government than they would in a unitary system.

III. Migration and the Vote for National Laws

When voters can move between jurisdictions in response to public policy changes, the vote for national preemptive laws will change. If migration is costless and if legal regimes are the only determinant of location, then, given G, national electoral support would fall in a hierarchical federal system. Voters are less likely to want national uniformity when they can move to congenial states.

Reality, of course, is somewhere between the extremes of a perfectly fixed population and a perfectly mobile one. Furthermore, different types of people have different moving costs, and the relation between these costs will help determine the political power of various groups at the state level. Thus, mobile groups may impose costs on immobile groups as a price for not leaving the state. Alternatively, undesirable groups may be induced to migrate by taxing or regulating them or by subsidizing their moving expenses. Variations in the cost of migration change the strategic environment. Those who can move cheaply have an advantage over those who cannot. Therefore, national decisions which require uniformity will be strongly supported by those who lose from the migration of others and opposed by those whose mobility permits them to gain at the expense of less mobile voters.

The possibility of migration has implications for national programs which seek to redistribute income from high-income to low-income families. Economic analyses of multiple government systems (e.g., Olson 1969; Oates 1972) conclude that interjurisdictional mobility

[8] It also, of course, includes those who are indifferent between any of these alternatives. Only people who are indifferent to all three alternatives are in both A_2 and A_3 or in both B_2 and B_3.

makes serious redistribution impossible at the state level. Progressive tax and spending policies must be carried out by preemptive laws at the national level. Interstate migration, however, also lowers the prospects for passage of national redistributive laws. If voters are mobile, support for a national preemptive law which redistributes income from the rich to the poor may be less than the support such legislation would obtain in a unitary system or in a federal system with no migration. Many who would support a particular progressive tax and spending program in a unitary system may oppose it in hierarchical federalism if the law preempts state "beggar-my-neighbor" laws which benefit these citizens. Selfishness will dominate altruism if the opportunity cost of altruism is too high.

IV. Does Federalism Matter in the Long Run?

The result in (3) applies when the number of x-type and y-type states is held constant (i.e., for a given G). One student of federalism (Riker 1975), however, argues that in the long run a federal system and a unitary one will converge to the same pattern of national laws. The view that federalism does not matter in the long run is based either on a hypothesis about how tastes change over time[9] or on a model that makes particular assumptions about the way votes change in response to new information and new circumstances (i.e., changes in G). The cases where federalism does not matter are contingent upon these assumptions. This section shows that it is possible to impose other plausible conditions on tastes, opportunities, and government structure and produce an equilibrium where x- and y-type states coexist in spite of majority support for x over y.

My analysis of state legislative choices challenges the idea that states are "lighthouses" that show the way to the federal government by enacting innovative laws. In my model, a law may spread to many states, not because it has been "tested" and found useful,[10] but rather

[9] Riker's argument turns on the overriding importance of individual tastes in determining political outcomes. He uses the example of white racism. "As long as whites strongly prefer racist institutions, one can expect institutions to be racist regardless of whether the country is federal or unitary. But when the preference for racist institutions weakens, then federalism helps racism by rendering difficult the enforcement of an anti-racist policy on the minority of white racists. So we can say that the beneficiaries of federalism get only marginal benefits on policy, but marginal or not, they are undoubtedly real" (1975, p. 156).

[10] See Walker (1969, 1971, 1973). By his choice of words, Walker appears to believe that early adopters of laws are progressive. He calls them "pioneers" (1969, p. 881). A glance at Walker's list of laws should warn anyone against this inference. Many involve the licensing of occupations such as barbers or real estate brokers. Others simply mandate the establishment of state agencies, many of which are required as a condition for receiving federal grants. For a different perspective on the same issue which emphasizes politicians' incentives to take risks, see Rose-Ackerman (1980).

because voters in y-type states want to avoid damage at the expense of others. Similarly, a national law may eventually pass in a federal system, not because a new initiative has been tested in the states, but rather because voters want to override the costs of spillovers and inconsistent laws.

Since a fully general analysis would be difficult to interpret, I concentrate on three special cases that capture the essential features of many actual situations. In these examples, I assume that tastes do not change over time, that voters have similar initial information, and that benefits that spill over to people in one group of states are costs to people in the other. In case 1, voters in x-type states receive positive spill-ins from out-of-state residents, and voters in y-type states bear costs. In case 2, voters in x-type states receive negative spill-ins from y-type states, and voters in y-type states benefit at the others' expense. Finally, in case 3, voters' beliefs in "states' rights" may conflict with their substantive position on other issues.

Case 1: Positive Spill-ins to x-Type States

Examples of case 1 are casino gambling or state-run lotteries in states where voters are concerned only with tax revenues and jobs. Taxes imposed on products sold to out-of-state consumers also fall into case 1. In these situations, A_2 and B_3 are empty sets.[11] Thus (3) becomes $N(A_3) \gtreqless N(B_2)$.

Since an x-type law leads to positive spill-ins, it is plausible to suppose that over time more and more states pass x-type laws (i.e., the set G changes). A few states institute lotteries, for example, and eventually many more follow their example. If this happens, people in set B now join set A. Many who favored a national lottery (i.e., those in B_1 and B_2) may now oppose it as a way of preserving their newly acquired positive spillovers (i.e., they move to A_3 and A_4, respectively).

As more states pass x-type laws, however, the level of net positive spillovers to each x-type state falls. In the case of casino gambling and lotteries, a state's consumers are less likely to gamble in other states if their own state has an x-type law.[12] In addition, negative spillovers to y-type states increase as more and more states export costs to them. The decline in net benefits and the increase in net costs as G changes

[11] Voters in A_2 favor preemption but oppose x in a unitary system. Voters in B_3 oppose preemption but favor policy x in a unitary system. Both of these preference patterns seem to be implausible if x-type states benefit at the expense of y-type states. This case would not hold if some people think that gambling brings costs (i.e., crowds, corruption).

[12] Similarly, tax incentives to encourage industry to locate in a jurisdiction provide few benefits to taxpayers if most states provide equivalent incentives. For an attempt to show the costs of interstate competition for business investment, see Jacobs (1979).

will raise the vote for a national preemptive law in individual x- and y-type states. Thus, a federal structure generates two influences that operate in opposite directions. A person whose state institutes a lottery is more likely to oppose federal preemption. However, as the number of states with lotteries increases, the benefits of multiplicity fall. If the first factor dominates the second, federalism will continue to matter even in the long run.

Case 2: Positive Spillovers to y-Type States

Examples of case 2 are minimum wage laws[13] or fair labor statutes that increase job opportunities in states without such laws and state pollution control laws where water and air cross state lines. Since in case 2 federalism permits y-type states to gain at the expense of x-type states, A_3 and B_2 are implausible preference patterns. Thus (3) becomes $N(B_3) \gtreqless N(A_2)$.

The majority of voters in x-type states have decided that, in spite of the costs, the law is worth having. Suppose that as time passes, voters in y-type states learn from the experience of those who have x-type (e.g., minimum wage) laws. If the benefits are greater than expected, then eventually more states pass minimum wage laws. People whose states enact such laws are now more likely to favor federal preemption. They no longer obtain the benefits of spillovers from other states, and they might like to prevent other states from benefiting at their expense. However, as the number of x-type states increases (G changes), the costs imposed on each x-type state fall and the benefits to each y-type state rise. This shift in costs and benefits should reduce support for a preemptive law in individual x- and y-type states. If the second (change in G) effect outweighs the first (change from B to A) effect, a national minimum wage law may not pass in a federal system even though many states have passed their own laws and $N(xRy) > N(yRx)$. Historical research, however, has stressed the cases where state initiatives led eventually to a federal statute,[14] leaving unanalyzed the many situations where the adoption of a law by many states has not been followed by a preemptive federal initiative.[15]

[13] Riker (1964, p. 146) uses the example of minimum wage laws that differ across states. "There is then much likelihood of capital flow from the high-wage localities to the low-wage localities for all those industries in which labor represents a high proportion of the cost. Aside from the imposition of a nationally uniform minimum wage, the only way that high-wage localities may counter this capital flow is by reducing the minimum wage level."

[14] See Fine (1956, pp. 353–85), who discusses the examples of antitrust laws, pure food and drug laws, protective labor legislation, social security, and welfare.

[15] Many of the laws studied by Walker (1969) have spread across many states but have never been adopted nationally. It is possible that some of these statutes might have majority support in a unitary system.

Case 3: States' Rights

In this case, some people put their belief in states' rights above their position on x (e.g., the abolition of slavery) and are in A_3 and B_3. Thus, no one favors imposing x on all states if they also oppose x in a unitary system (i.e., A_2 and B_2 are empty sets). Therefore, (3) becomes $N(A_3 \cup B_3) > 0$.

The vote for x in a federal system is certain to be less than the vote in a unitary system, and federalism may continue to matter so long as preferences remain constant.[16] Assume, however, that people with a substantive preference for x over y favor states' rights only if the results of independent state choices are not too far away from their position on x. Then, it is easy to tell a simple story where an unanticipated change leads to convergence of a federal and a unitary system. If the number of y-type states (e.g., slave states) increases exogenously, people who are opposed to slavery may now choose federal preemption over states' rights (i.e., some of those in A_3 and B_3 move to A_1 and B_1). Therefore, x may now pass at the national level. An exogenous change that seemed to favor y leads in fact to its repudiation by the nation. Thus, it may not be necessary to assume a growth in Northern antislavery sentiment to explain the evolution of pre–Civil War politics. As slavery spread across the South after the invention of the cotton gin, even people whose antipathy to slavery was constant might have supported a national policy of abolition if they thought that more and more states would permit slavery (i.e., they might shift from

[16] For symmetry, we could include a fourth case that appears to be of less empirical importance. In this final case, people believe in uniformity or in a strong central government and are willing to support a national law even when they would oppose the law in a unitary system. Thus, A_3 and B_3 are empty sets, (3) becomes $0 < N(A_2 \cup B_2)$, and the vote for x is always larger in a federal system than in a unitary system. If y is a status quo which permits the private market to operate with a minimum of government interference, then case 4 is illustrated by business managers who favor laissez-faire (yRx) but if faced with a mixture of differing state laws would rather have a uniform federal regulatory statute because it permits them to reduce costs. Employers who sell their products in many states often fit this preference pattern. They would rather not be regulated at all, but being regulated in some states is worse than a uniform national standard. For example, sellers of bottled mineral water are beginning to face a variety of labeling laws enacted by different states. According to *Business Week*: "Most bottlers say they would not oppose reasonable, uniform regulations. . . . Industry executives shudder at the prospect of trying to comply with rules that could vary widely among states" ("Mineral Water Could Drown in Regulation," *Business Week* [June 11, 1979]). State trucking regulations are sometimes inconsistent and costly for truckers. Thus, state regulations of the mudguards required on interstate trucks were inconsistent, and an Illinois law was overturned by the Supreme Court as interfering with interstate commerce (*Bibb v. Navajo Freight Lines, Inc.*, 359 U.S. 520 [1959], discussed in Tribe [1978, p. 339]). Similarly, the Supreme Court ruled that Wisconsin's prohibition against twin trailers on interstate highways was an unconstitutional interference with interstate commerce ("Breaking a Bottleneck in Long-Haul Trucking," *Business Week* [March 7, 1978]).

A_3 to A_1). When slavery looked as if it might expand into the western states, people who had favored a federal solution might begin to support abolition.[17]

V. Conclusions

Political economists have generally recognized that realistic political systems must include some interjurisdictional spillovers as a cost of providing citizens with a choice of public service levels. Intergovernmental grants may reduce these interjurisdictional costs (Breton 1965) but will not eliminate them. Because both the dispersion of tastes across the population and the level of externalities differ for different public services, some analysts have argued that a federal system is the best way to accommodate these conflicting tendencies (e.g., Oates 1972). In making this recommendation, however, these authors fail to make clear an important idealization inherent in their analysis. They assume that it is possible to assign functions unambiguously to levels of governments so that constitutional structure only has an impact on spillovers and on the position of minorities. In fact, it will often be impossible to assign responsibilities neatly to a particular political level. Indeed, this overlap is the characteristic feature of contemporary "cooperative" federalism. Given this fact, the present paper has shown that, when authority is divided, the choices of lower-level governments can have important consequences for the decisions of higher-level governments. Even when the central government has the power to preempt state and local laws, its democratic choices will depend upon the strategic position of citizens living under alternative state legal regimes. The essential difference between a hierarchical federalism and a unitary system is the difference in the status quo. This difference affects the vote on national legislation and the bargaining power of individuals. Individuals may vote against the extension of a state law to the nation as a whole even though they

[17] Potter (1976), in a history of the period from 1848 to 1861, stresses the difficult trade-offs faced by many people. He writes that "the problem for Americans, who, in the age of Lincoln, wanted slaves to be free was not simply that Southerners wanted the opposite, but that they themselves cherished a conflicting value: they wanted the Constitution, which protected slavery to be honored, and the Union, which was a fellowship with slaveholders to be preserved" (pp. 44–45). The Northern public "placed their antislavery feelings in a context of state action, accepting personal responsibility for slavery within their own particular states" (p. 46). In contrasting the position of Stephen Douglas in 1854 with that of his opponents, Potter shows how these trade-offs were resolved in different people. "Douglas was a vigorous believer in the democratic principle of local autonomy, but his opponents were equally vigorous believers in the moral primacy of freedom. . . . Douglas cared more about the Union than about the eradication of slavery and would never push the slavery issue to a point where it imposed too much strain upon the Union. Many antislavery men thought the Union hardly worth preserving so long as it had slavery in it" (pp. 172–73).

would favor the law in a system with only one government. They may wish to extend a state law to all citizens although they would oppose the law in a unitary system.

The analysis suggests a promising area for future empirical research. Congressional votes on particular issues might be associated with existing state laws,[18] and the timing of federal passage of laws in several areas could be related to the patterns of state adoptions of similar laws. Empirical work is also needed on specific policy areas. For example, one might study the impact of local power over schools on state and federal education policy or see how federal efforts to reduce poverty have been conditioned by prior state and local efforts.

My discussion of federalism also suggests that, at the point of constitutional choice, people are more likely to support a federal system with strong lower-level governments, the fewer the strategic possibilities open to individual states. Strong low-level governments will be attractive if people are grouped geographically by their taste for public services and if it is difficult to impose costs on other jurisdictions. This implies that many people would support constitutional constraints on state governments that limit the states' strategic behavior while retaining many of the benefits of variety and experimentation. I would also expect that, if interjurisdictional migration is possible, then those who favor aid to currently disadvantaged groups at the expense of mobile capital and labor resources will favor a strong central government.[19] It is a commonplace in economic analyses of federalism to note that low-level governments cannot carry out redistributive policies (e.g., Oates 1972). My point, however, goes beyond this conclusion to the observation that if states try to gain at the expense of other states, then interstate redistribution can occur that bears little relation to anyone's notion of social justice and may end up making all households worse off. The end result depends upon each state's strategic position, that is, its ability to export taxes and import benefits.

References

Breton, Albert. "A Theory of Government Grants." *Canadian J. Econ. and Polit. Sci.* 31 (May 1965): 175–87.

[18] McLure (1967, p. 72), in a study of the export of state taxes, recognizes that "states with the largest export rates can be expected to favor state or local assumption of governmental activities while those with low export rates might reasonably favor federal action."

[19] Thus, Dye (1973, p. 62) writes that "urban interests, low-income groups, blacks, ethnic groups and labor organizations frequently turn to the national government for help. States' rights arguments have little appeal to those groups, which are important in the national electorate, but do not constitute majorities in the large number of sparsely settled states."

Cover, Robert M., and Aleinikoff, T. Alexander. "Dialectical Federalism: Habeas Corpus and the Court." *Yale Law J.* 86 (May 1977): 1035–1102.

Dye, Thomas R. *Politics in States and Communities.* 2d ed. Englewood Cliffs, N.J.: Prentice-Hall, 1973.

Elazar, Daniel J. *The American Partnership: Intergovernmental Co-operation in the Nineteenth-Century United States.* Chicago: Univ. Chicago Press, 1962.

Fine, Sidney. *Laissez Faire and the General-Welfare State.* Ann Arbor: Univ. Michigan Press, 1956.

Grodzins, Morton. "The Federal System." In *President's Commission on National Goals for Americans.* Englewood Cliffs, N.J.: Prentice-Hall, 1960.

———. *The American System. A New View of Government in the United States*, edited by Daniel J. Elazar. Chicago: Rand McNally, 1966.

Jacobs, Jerry. "Bidding for Business. Corporate Auctions and the 50 Disunited States." Mimeographed. Washington: Public Interest Res. Group, August 1979.

McLure, Charles E., Jr. "The Interstate Exporting of State and Local Taxes: Estimates for 1962." *National Tax J.* 20 (March 1967): 49–77.

Oates, Wallace E. *Fiscal Federalism.* New York: Harcourt Brace Jovanovich, 1972.

Olson, Mancur, Jr. "The Principle of 'Fiscal Equivalence': The Division of Responsibilities among Different Levels of Government." *A.E.R. Papers and Proc.* 59 (May 1969): 479–87.

Posner, Richard A. *Economic Analysis of Law.* 2d ed. Boston: Little, Brown, 1977.

Potter, David M. *The Impending Crisis, 1848–1861.* New York: Harper & Row, 1976.

Riker, William H. *Federalism: Origin, Operation, Significance.* Boston: Little, Brown, 1964.

——— "Federalism." In *Handbook of Political Science.* Vol. 5, *Governmental Institutions and Processes*, edited by Fred I. Greenstein and Nelson W. Polsby. Reading, Mass.: Addison-Wesley, 1975.

Rose-Ackerman, Susan. "Risktaking and Reelection: Does Federalism Promote Innovation?" *J. Legal Studies* 9 (June 1980): 593–616.

Rothenberg, Jerome. "Local Decentralization and the Theory of Optimal Government." In *The Analysis of Public Output*, edited by Julius Margolis. New York: Columbia Univ. Press (for Nat. Bur. Econ. Res.), 1970.

Sen, Amartya K. *Collective Choice and Social Welfare.* San Francisco: Holden Day, 1970.

Tribe, Laurence H. *American Constitutional Law.* Mineola, N.Y.: Foundation Press, 1978.

Tullock, Gordon. "Federalism: Problems of Scale." *Public Choice* 6 (Spring 1969): 19–29.

Walker, Jack L. "The Diffusion of Innovations among the American States." *American Polit. Sci. Rev.* 63 (September 1969): 880–99.

———. "Innovation in State Politics." In *Politics in the American States: A Comparative Analysis*, 2d ed., edited by Herbert Jacob and Kenneth N. Vines. Boston: Little, Brown, 1971.

———. "Comment. Problems in Research on the Diffusion of Policy Innovations." *American Polit. Sci. Rev.* 67 (December 1973): 1186–91.

Wechsler, Herbert. "Political Safeguards of Federalism." In *Selected Essays on Constitutional Law 1938–1962*, edited by Edward L. Barrett, Jr., et al. St. Paul, Minn.: West Publishing (for Assoc. American Law Schools), 1963.

Wheare, Kenneth C. *Federal Government.* 3d ed. London: Oxford Univ. Press, 1953.

Winter, Ralph K., Jr. "State Law, Shareholder Protection, and the Theory of the Corporation." *J. Legal Studies* 6 (June 1977): 251–92.

Ylvisaker, Paul. "Some Criteria for a 'Proper' Areal Division of Governmental Powers." In *Area and Power: A Theory of Local Government*, edited by Arthur Maass. Glencoe, Ill.: Free Press, 1959.

Market Provision of Price-excludable Public Goods: A General Analysis

Michael E. Burns

Monash University, Melbourne

Cliff Walsh

University of Adelaide

The *demand distribution* is employed as a novel and powerful basis for analyzing monopoly provision of price-excludable public goods (and could be used to analyze other market structures or, in some circumstances, private goods provision). For (uniform) per-unit, all-or-none, two-part, and multipart pricing, we identify: characteristics of revenue functions, relative profitability, and operational procedures for selecting price-output levels. Under all these strategies, rationing of some consumers *by output* is required, and a positive price-output relationship may arise. In general the revenue ranking of pricing strategies is sensitive to the distribution of demand and, though multipart pricing can be expected to be revenue dominant, uniform per-unit pricing emerges as a surprisingly robust strategy even before operational complexity is fully taken into account.

Recent contributions to the economics literature reflect a growing interest in private provision of commodities which possess one

Work on this paper was completed while both authors were in the Department of Economics, Monash University. The notion of the demand distribution, to our knowledge, was first suggested in earlier unpublished work of the first-mentioned author. We have benefited from comments of participants in a Monash workshop, a Virginia Polytechnic Institute seminar, and the 1978 Public Choice meetings. John Head, Yew-Kwang Ng, Geoffrey Brennan, and David Friedman provided useful observations, but our greatest debt is to William Oakland, whose extensive and perceptive comments have been of considerable assistance. The authors nonetheless accept the conventional attribution of errors to themselves.

[*Journal of Political Economy*, 1981, vol. 89, no. 1]
© 1981 by The University of Chicago 0022-3808/81/8901-0004$01.50

characteristic of a public good ("jointness," or "nonrivalness in consumption") but not the other ("impossibility of exclusion"): Exclusion is assumed to be commercially feasible, so that anyone unwilling to pay the required price can be denied access to any or all of the units produced.[1] This important class of commodities can be termed *price-excludable public goods* or (reflecting a characteristic they retain rather than one they lose) *joint goods*. Television and radio transmissions and the services provided by transportation facilities (rivers, roads, airports, ships, buses, and airplanes), entertainment facilities (theaters, galleries, museums), recreational facilities (national parks, football stadiums), and information serve as examples, of varying degrees of purity:[2] All are important in modern societies, and virtually all have been suspected of being suboptimally provided by the private sector.

Competitive models dominate the published literature, and Samuelson's (1954, 1955) seminal analysis of pure public goods has clearly influenced the questions asked and the methodology employed. Analysis of the behavior of and interaction between individual profit-seeking entrepreneurs has been neglected in favor of somewhat speculative predictions of the likely characteristics of market equilibrium so that efficiency issues can be immediately examined. This is true, to some extent, even of the important contribution of Oakland (1974), in which it is suggested that competition will generate an equilibrium with unfamiliar characteristics in the joint goods context, with consumers facing an increasing step function of prices for homogeneous consumption units.[3]

In contrast, we present an analysis of joint goods provision with the individual profit-maximizing entrepreneur as its focal point. However, we depart from the analogous private goods methodology in one important respect: The familiar notion of the aggregate demand curve is replaced by the *demand distribution* as a novel, powerful, and operationally meaningful building block, reflecting the fact that the number of consumers of particular output units plays a critical role in joint goods analysis even under simple pricing strategies. Using this

[1] See, in particular, Buchanan (1967), Thompson (1968), Demsetz (1970), Oakland (1974), Auster (1977), Lee (1977), and Brennan and Walsh (1979); and also working papers by Brennan and Walsh (1978) and Burns (1979).

[2] Pure joint goods (like pure public goods) would permit an indefinitely large number of consumers. The examples cited (except perhaps information) fall well short of this but, as long as they remain uncongested, can still be analyzed as if they were "pure" (like "local" public goods). By treating costs as varying with intensity of consumption, our approach could also be applied to the impure cases.

[3] We accept Oakland's description of equilibrium, but his analysis of its attainment is essentially circular. The process of establishment of equilibrium must be substantially different from that in the private goods case. In fact, it may be that the process is analogous to the seriatim introduction of "new" goods

basis, our central contribution is a general exposition of the neglected case of monopoly provision of joint goods, and for a variety of pricing strategies we are able to identify characteristics of revenue functions, relative profitability, and operational procedures for selecting price and output levels. Nonetheless, our framework can also provide a basis for analysis of competitive joint goods provision and for analyzing marketing strategies for *private goods*, particularly where constant marginal costs prevail.

In Section I we introduce the demand distribution, provide a geometric interpretation of it, and set out basic assumptions and definitions. Sections II and III examine the characteristics of a range of feasible marketing strategies (applied uniformly across individuals) from per-unit pricing to sophisticated multipart pricing procedures, each of which has been discussed in the private goods literature and has a basis in common business practice.[4] Section IV contains an illustrative analysis using a specific distributional form, while Section V brings together the main results with final observations.

I. Framework of the Analysis

For monopoly provision of private goods under uniform per-unit pricing, information about *aggregate* demand provides an adequate basis for decision making. Since each output unit is exclusively consumed by a single individual, the profit-maximizing price will never be less than marginal production cost, and hence the firm will always produce as many units as are demanded by all individuals at that price. The aggregate demand curve thus completely specifies the relationship between per-unit price and *both* aggregate consumption *and* production, directly determining the revenue-output relationship. However, even for a private good producer, the information revealed by the aggregate demand curve is not sufficient when discriminatory pricing schemes are being considered: for these pricing strategies, information is required on the "composition" of demand for the product.

In contrast, for monopoly production of price-excludable public goods, information on aggregate demand is inadequate even under uniform per-unit pricing. Since each production unit can be fully and equally consumed by all individuals, output need never exceed that

[4] Dupuit ([1844] 1952) was an acknowledged authority on price discrimination and influenced the adoption, by the nineteenth-century French public sector, of quite sophisticated pricing procedures. More recent literature includes general analyses by Gabor (1955–56), Puu (1964), Ng and Weisser (1974), and Yamey (1974), and contributions on quantity discounting by Buchanan (1952–53) and Berglund (1964). In many cases the analysis has been applied to goods with significant jointness elements.

required to satisfy the highest demand individual at any price. Moreover, the per-unit price faced by each individual can be less than marginal production cost since many units are jointly consumed, but this would necessitate the rationing of some high-demand individuals by output rather than by price. Consequently, not only does the conventional aggregate demand curve not define the relationship between price and output for joint goods, in general it need not even define the relationship between price and aggregate consumption. Operationally, given his knowledge of production costs, the producer of a joint good will be concerned to identify the maximum revenue obtainable from (various) given output levels, and this critically depends on the composition of demand. Specifically, he will be interested in the number of individuals who would purchase (at least) a certain quantity when confronted with a particular (revenue-maximizing) price, since this determines his marginal revenue. Compared with his private goods counterpart (who can work simply with output and price), the joint good producer must consider output, revenue-maximizing price, *and* the number of consumers of the marginal production unit. Accordingly, we have adopted as the fundamental building block of our analysis a construct which takes account of the critical distributional characteristics of demand, is based on variables that may be empirically observable in the normal process of operation, and has a basis in conventional demand theory. This construct we term the "distribution of demand" or, more succinctly, the *demand distribution*.

The Nature of the Demand Distribution

A point on the surface of the demand distribution defines the number of individuals, n, who would each consume at least q units of output if the joint good was made available at a per-unit price of p.[5] Formally (adopting the conventional private good assumptions that we are dealing with large numbers of individuals and that prices and quantities are divisible) we assume that the distribution may be represented by a well-defined function, $n = n(p,q)$, with the properties

$$\frac{\partial n}{\partial p}, \frac{\partial n}{\partial q} < 0, \tag{1}$$

so that, by the implicit function theorem, the inverse function, $p = p(n,q)$, is also well defined.

It is appropriate that we should briefly discuss the basis for these

[5] Note esp. that n does not refer to the number who would consume some part of output q, but rather to those who would consume it all.

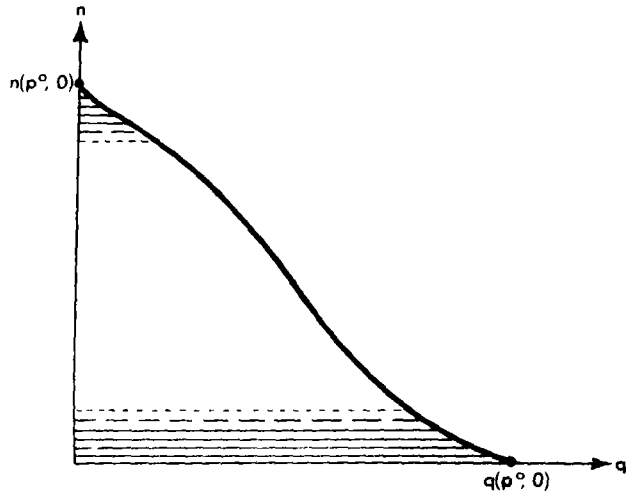


FIG. 1.—Cross section through demand distribution with price constant at p^0

specific assumptions. In figure 1 we show, for a particular per-unit price, how the number of individuals consuming each unit of output is functionally determined by the distribution of demand. For n here to be a well-defined function of q clearly we must have the demands of individuals arranged in order of magnitude; but, more than this, we require that there never be a significant number of individuals with identical demands (at any uniform price). Thus, not only have we ruled out the possibility of a significant number of identical individual demand curves but also (perhaps more interestingly), while accepting that many such curves may intersect, we have ruled out the possibility that a significant number will intersect at identical values of q (and p). The strict inequality assumed in equation (1) also involves further considerations. First, it implies that variables which influence the pattern of demand (e.g., tastes, income, and wealth) are assumed to be distributed continuously among the population. Were this not the case, significant changes in q could be associated with zero changes in n and a well-defined inverse function, $p(n,q)$, may not exist. Second, besides these implications regarding the relationship between n and q (with p constant), equation (1) also implies the assumption that individual demand curves (defined as $q_i = q_i[p]$) are downward sloping.[6]

[6] William Oakland has suggested a neat demonstration of how the distributional function, $n(p,q)$, might be derived formally from individual demand behavior. Let the inverse demand relations of a subset of individuals be given by $p = p(q,a)$, with the shift parameter distributed according to $f(a)$ over $(a,\bar{a})$. Then, for given p and q, n is determined by $n(a) = \int_a^{\bar{a}} f(s)ds$. Moreover, the intersecting case is accommodated immediately by considering any number of different classes of such demand relations. E.g., with two classes, $p = p(q,a)$ and $p = \phi(q,b)$, we would have separate density functions for a and b and define $n = n(a) + n(b)$.

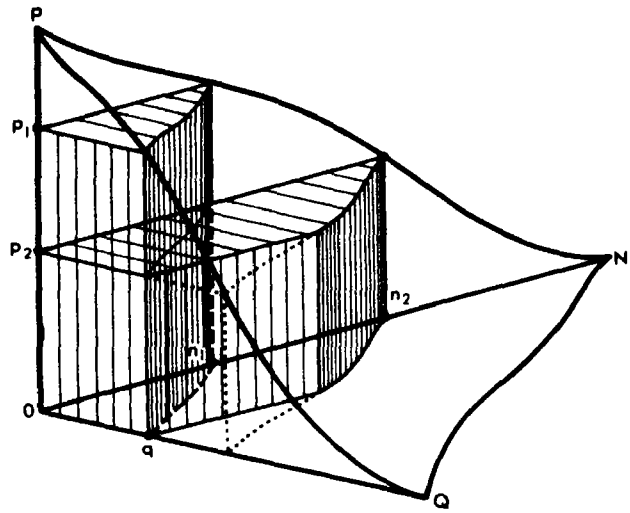


FIG. 2.—Demand distribution

The sum of all these restrictions yields the demand distribution as defined above, a geometric representation of which is provided in figure 2. Here the boundary values P, Q, and N define, respectively, the maximum price that would be paid for the initial unit of output, the maximum quantity that would be demanded by any individual at zero price, and the total number of individuals. Also shown are the solids whose volumes indicate the revenues obtained from alternative per-unit pricing arrangements associated with a particular output level. These revenue volumes reflect the fact that, for a joint good, each unit produced may be consumed by all individuals willing to pay the price, p, and are especially relevant to the analysis undertaken in Section II.

The All-or-Nothing Distribution

The incorporation into our analysis of the (desirable) feature that individual demand curves may intersect has important consequences for the treatment of all-or-nothing procedures. With respect to the perspective provided in figure 1, the presence of such intersections means that the order in which individuals' demand curves are arranged must vary according to the (constant) price level considered if ordering with respect to magnitude is to be preserved. Thus, in general it may *not* be assumed that a cross section through the distribution with n held constant will indicate the demand relation of a particular individual (although this will be so in the nonintersecting case). Even if we make the conventional assumption of zero income

effects so that $p(n,q)$ indicates the marginal evaluation of different consumption units, the integral of this inverse function, given n, cannot be assumed to measure the "willingness-to-pay" of a particular individual. For this reason, and because users of all-or-nothing procedures will generate market observations on variables different from those relevant to the per-unit pricing case, the analysis of these strategies (presented in Section III) has been undertaken with respect to an alternative (but related) distributional construct, the *all-or-nothing distribution*. A point on the surface of this distribution defines the number of individuals, n, who would purchase a bundle of q units of the good at an average (per-unit) charge C rather than go without the good altogether. We assume, analogously to the demand distribution, that the all-or-nothing distribution may be represented by a well-defined function, $n = n(C,q)$, with the properties

$$\frac{\partial n}{\partial C}, \frac{\partial n}{\partial q} < 0, \tag{2}$$

again ensuring that the inverse function, $C = C(n,q)$, is well defined.

Some Useful Definitions

Finally, later discussion is considerably assisted by reference to particular concepts which may be formally defined at this stage. The definitions made here with respect to the demand distribution may also be applied to the all-or-nothing distribution.

DEFINITION 1: A level of output q will be termed *efficient with respect to (a particular pricing procedure)* if under that procedure: $R(q) > R(q^*)$ $\forall$ $q^* < q$ (where R denotes total revenue).

DEFINITION 2: A *regular distribution of demand* has the property: $\partial^2 R(q_0)/\partial p^2 < 0$ (where $R[q_0] = p[n,q_0] \cdot n[p,q_0]$).

DEFINITION 3: A *weakly restricted distribution of demand* has the property that: $\partial p^*(q_0)/\partial q_0 < 0$ (where $p^*[q_0]$ is that price which maximizes $R[q_0]$).

Definitions 2 and 3 are restrictions on the nature of the demand distribution, the former requiring concavity of the revenue function relating to consumption solely in respect of the q_0th unit of output.[7] Some of the more complex operational results derived in the paper require these plausible restrictions to hold, and in these cases the

[7] Definition 2 reflects the conventional simplification adopted in private good analysis that marginal revenue functions are strictly downward sloping. The restriction implied by definition 3 is lent plausibility by the observations that it will always be true if cross sections through the distribution for q constant are (mathematically) similar and that even where this is not so the overall trend in $p^*(q)$ must be negative.

results have been presented as propositions with proofs clearly identified in the text.

II. Uniform Per-Unit Pricing

We have previously emphasized the important distinction between output and aggregate consumption in the joint goods context. Since, operationally, the entrepreneur will be concerned with the profitability of various levels of *output* and because his production cost conditions need have no unusual features, our analysis of alternative marketing strategies can focus on his search for the revenue-maximizing price (and/or charge) associated with each potential output level and, more generally, on the characteristics of the revenue functions he faces. Initially, we suppose that the joint good producer seeks to maximize profit while setting a price which is uniform across both individuals and output units.

Basic Characteristics of the Uniform Per-Unit Price

The unusual nature of his pricing decision is made transparent by first supposing that the entrepreneur behaves analogously to his private good counterpart, allowing all individuals to consume as much as they wish at the prevailing per unit price. Since, for joint goods, output would be then determined by the highest individual consumption level demanded at each price, this strategy can be alternatively represented by the assumption that at any output level the producer selects the maximum price at which anyone is willing to consume the entire output. This procedure we call *Maximum Uniform Pricing* (MUP): In figure 2, for output q it involves setting price p_1, at which price the highest demand individual consumes the entire output, and a total of n_1 individuals consume at least part of it.

The revenue yielded by this pricing procedure, for any output level q_0, is given by:

$$R(q_0)/\text{MUP} = \int_0^{q_0} p_0 \cdot n(p_0,q)dq, \tag{3}$$

where $p_0 = p(0,q_0)$ and is constant throughout the integration. Differentiation of (3) with respect to q_0 yields the marginal revenue (MR) function:[8]

$$\begin{aligned}\text{MR}(q_0)/\text{MUP} = {} & \frac{\partial p_0}{\partial q_0}\int_0^{q_0} n(p_0,q)dp \\ & + p_0\int_0^{q_0} \frac{\partial}{\partial p} n(p_0,q)\,\frac{\partial p}{\partial q}\,dq + p_0 \cdot n(p_0,q_0).\end{aligned} \tag{4}$$

[8] Discussion of the appropriate algebraic methods for derivation of this expression is to be found in Courant (1960), p. 220.

Note, however, that the last term in this expression, representing the effect of a change in output, price held constant, in the limit approaches zero because $n(p_0,q_0)$ itself tends to zero. Thus, $\mathrm{MR}(q_0)/\mathrm{MUP}$ can be written as:

$$\mathrm{MR}(q_0)/\mathrm{MUP} = \frac{\partial}{\partial p}[R(q_0)/\mathrm{MUP}]\frac{\partial p}{\partial q}. \tag{4a}$$

This marginal revenue function has a number of interesting properties (identified in the illustrated example in fig. 4 below), but more importantly it immediately suggests that for joint goods the optimal (i.e., revenue-maximizing) pricing strategy differs from MUP (the private goods analogue). Any level of output which is *efficient with respect to MUP* must have $\mathrm{MR}(q)/\mathrm{MUP} > 0$, and it follows directly from (4) and (4a) that, at any such output, revenue would be increased by lowering price (with output held constant). This is clearly illustrated in figure 2, where for output q, lowering price from p_1 (i.e., $p[q]/\mathrm{MUP}$) to p_2 increases total revenue: revenue declines from the original consumption units, but this is more than offset by the fact that each output unit is now more intensively consumed. Thus, *the Optimal Uniform Price, $p(q)/OUP$, will be less than the Maximum Uniform Price, $p(q)/MUP$, at any output level where marginal cost is positive.* Moreover, *some individuals* (all those for whom $q_i[p] > q$) *will be rationed by output while others* (for whom $q_i[p] \leq q$) *are rationed by price*, and this is indeed a general characteristic of optimal pricing for joint goods which does not apply for private goods.[9]

Some General Properties of Revenue Functions under OUP

Some insights into more general properties of the relevant revenue functions are straightforward. If OUP is charged for an output level q_0, we must have

$$\frac{\partial[R(q_0)/\mathrm{OUP}]}{\partial p} = 0. \tag{5}$$

It follows trivially, either from arguments similar to (3) and (4) above or from geometric considerations, that:

$$\mathrm{MR}(q_0)/\mathrm{OUP} = [p(q_0)/\mathrm{OUP}] \cdot n[p(q_0)/\mathrm{OUP},q_0]. \tag{6}$$

[9] The possible need to ration by output for per-unit pricing was first recognized in Brennan and Walsh (1978). Strictly speaking, rationing applies only if at the profit-maximizing output $\mathrm{MC} > 0$ (see subsection, "Some General Properties of Revenue Functions under OUP," following). Moreover, while for $p > \mathrm{MC} > 0$ rationing cannot sensibly be considered to apply, the range of output levels for which this occurs is negligible when large numbers of consumers are involved. It should be observed, also, that the use of continuous mathematical methods makes it appropriate to regard $n[p(q_0)/\mathrm{OUP},q_0] > 0$ as indicative of rationing even for the special case of $1 > n[p(q_0)/\mathrm{OUP},q_0] > 0$.

That is, *under OUP marginal revenue (with respect to output) will be equal to the revenue obtained solely in respect of the final (i.e., least consumed) unit of output.* Regarding the characteristics of the function MR(q)/OUP, it clearly can never exceed max $[p(n,q) \cdot n(p,q)]$ which, given the properties of the demand distribution (1) above, must decrease as output increases. Since MR(0)/OUP = max $[p(n,0) \cdot n(p,0)]$ marginal revenue under OUP must obtain its maximum value at zero output, although it may sometimes increase with output at some nonzero production levels.[10]

The results (5) and (6) also yield insight into the characteristics of OUP when output is selected to yield the maximum possible revenue. Since this situation must involve $p(q)$/OUP > 0 but MR(q)/OUP = 0 it follows from (6) that there will be no rationing. In this case the OUP solution has MUP characteristics and, since (5) always holds for OUP, is in fact identical to the revenue-maximizing solution for MUP.

Output and Price Determination under OUP

Since profit-maximization requires the equality of marginal cost and marginal revenue, we have begun by deriving and identifying the properties of the marginal revenue function appropriate to OUP, properties which (with respect to output) turn out to be similar to those observed in the private good case. The actual process of output determination under OUP is not unduly complicated. Once the Optimal Uniform Price has been (experimentally) determined for a particular output level it follows directly from (6), providing consumption of the final unit of output is known or can be estimated, that marginal revenue is also determined. However, the decision to increase or decrease output must be accompanied by a decision regarding price adjustment, and here the user of OUP is confronted, in complete contrast to the conventional (private goods) wisdom, by the possibility that the profit-maximizing price may rise as the profit-maximizing output increases, ceteris paribus. In order to show this we first derive a general result:

PROPOSITION 1: The Optimal Uniform Price for a given output level, $p(q)$/OUP, may generally be expected to lie in the range bounded by the previous such price, $p(q - dq)$/OUP, and the price which maximizes revenue solely in respect of the qth unit of output, $p^*(q)$. This will always be so if the demand distribution is *regular*.

PROOF: If a distribution is regular, for a given q and dq, the revenue

[10] The operationality of expression (6) does, of course, require that it is possible to identify how many individuals consume the entire output. Lack of this information in no way invalidates the results derived for total and marginal revenue functions throughout the paper, but it must be recognized that practical determination of MR often may be less straightforward than is implied by the simplicity of the stated results.

expression $p(n,q) \cdot n(p,q)dq$ is a concave function of price, as therefore must be the sum of any such expressions. The revenue obtained from per-unit pricing q units of output may be expressed as $\int_0^{(q-dq)} p(n,q) \cdot n(p,q)dq + p(n,q) \cdot n(p,q)dq$, which is therefore the sum of two concave functions of price. As such its maximum must be attained at a price bounded by those prices which maximize each of its parts, $p(q - dq)$/OUP and $p^*(q)$. Even if the functions are not globally concave with respect to price, since they are continuous they must be locally concave at their maxima. Where the maxima are not at significantly different prices, as well as in those cases where convexities in the revenue functions are not severe, we might also expect $p(q)$/OUP to be bounded as suggested. □

This conclusion, which is useful in the derivation of later results, also has some operational content. It suggests that in the absence of information about the trend of $p^*(q)$, larger output expansions could initially be accompanied by a price *decrease*, and for *small* increases in output (since $p^*[q + dq] \rightarrow p^*[q]$) perhaps price initially should be set between $p(q)$/OUP and $p^*(q)$. More immediately, however, it leads to a fundamental observation:

PROPOSITION 2: The Optimal Uniform Price, $p(q)$/OUP, may sometimes increase as output increases. However, this will never occur if the demand distribution is *regular* and *weakly restricted*.

PROOF: The first part follows directly from the possibility (illustrated in fig. 3, where $p^*[1]$ applies strictly to the initial unit of output and hence is equal to $p[1]$/OUP) that $p(q - dq)$/OUP may be less than

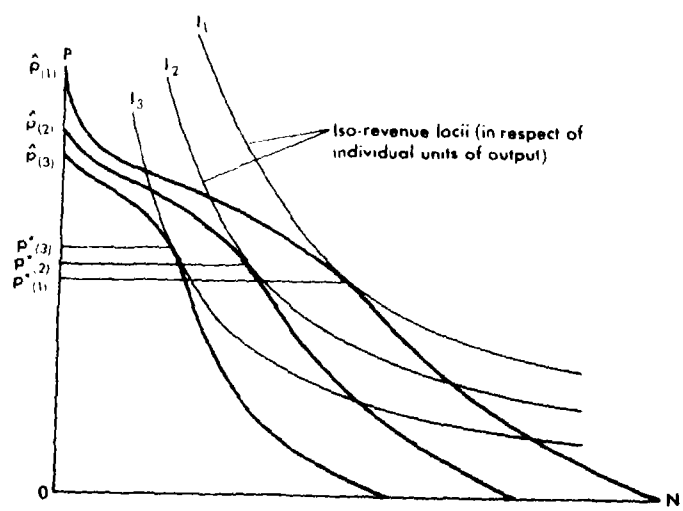


FIG. 3.—Distribution characteristics yielding price increases as output increases under OUP.

$p^*(q)$. For the second part, given a weakly restricted distribution, so that $p^*(q - dq) > p^*(q)$, and regularity, it follows from proposition I that

$$\begin{aligned} p(0)/\text{OUP} &= p^*(0) > p^*(dq) \\ \therefore p(0)/\text{OUP} &> p(dq)/\text{OUP} > p^*(dq) > p^*(2dq) \\ \therefore p(q - dq)/\text{OUP} &> p(q)/\text{OUP} > p^*(q). \ \square \end{aligned}$$

The phenomenon of (optimal) price rising when output is increased (i.e., for reductions in marginal cost) appears to be unique to the joint goods context. Moreover, even in the case of regular and weakly restricted distributions, the more conventional fall in price (as output rises) may be very small compared with that expected for a private good.[11] Other unusual phenomena, though perhaps to be given no more attention than peculiarities which can arise in traditional private goods cases, may also occur. For example, the optimal number of individuals to ration may (perversely) increase as output increases.

An Overview of Results for Uniform Per-Unit Pricing

Except in the special case where marginal cost is zero, Optimal Uniform Pricing of joint goods for any output involves a price lower than the Maximum Uniform Price for that output, and, consequently, some individuals will be rationed by output. Marginal revenue for OUP can be defined straightforwardly as the revenue obtained solely in respect of the last unit of output; the characteristics of the MR function are similar to those applying to private goods; and prescriptions regarding price and output determination can be obtained with only slightly more difficulty than in the private goods case. However, the Optimal Uniform Price may increase as output increases, and, even if this result does not arise, the price reduction associated with output increases may be small relative to those which would occur if the good was private.

III. More Complex Pricing Strategies

Observation of common business practice suggests that in the course of their marketing operations firms can generate information which

[11] Brennan and Walsh (1978) isolate a result in which price increases with output, but one which applies to specific points (actually "gaps") in the price-output relationship and which appears to emerge from the small-numbers case they deal with. Our analysis, however, using large numbers, confirms that the whole trend of prices may be upward over some output ranges. Our other result—that price falls may be small—is illustrated by the linear distribution example in Section IV since $P/2 > p(q)/\text{OUP} > P/3$, where P is the price intercept of the highest demand individual.

permits them to adopt pricing strategies more sophisticated than uniform per-unit pricing. Given the "service" nature of the typical examples, monopoly producers of joint goods may be well placed to consider the use of all-or-nothing charges, entrance or license fees, and procedures which price units (or bundles of units) separately, though on operational grounds we assume that such procedures are applied uniformly across consumers. As with per-unit pricing, our analysis of these possibilities is concerned to identify basic characteristics of the optimal pricing arrangement and operational aspects of price and output determination. We are now also concerned, however, to consider the relative profitability of the alternatives open to the firm, and for this purpose the demand and all-or-nothing distributions become all the more important. However, to enable clear-cut results to be generated we are obliged to assume that income effects are negligible. Since substitution effects can generally be expected to dominate income effects, it is likely that our results are broadly applicable, but the possibility of conclusions being reversed by income effects in some instances cannot be ruled out.

All-or-Nothing Pricing

All-or-nothing type pricing schemes are exemplified by "season ticket only" offers for a series of operas or Shakespeare plays and flat-rate bus or subway fares. We assume that the same all-or-nothing offer confronts all individuals for a particular output bundle offered by the producer, though it is convenient to express this offer in terms of an average (per-unit equivalent) *charge* which must be paid for all units in the bundle. The Optimal (revenue-maximizing) All-or-Nothing Charge (OAN) will be such that *the marginal consumer pays his average valuation for the units in the bundle and the OAN strategy inherently involves rationing by output.*

It is interesting to note that, in the special case where individual demand curves do not intersect, there is an obvious symmetry between the all-or-nothing and per-unit pricing procedures; as is discussed more fully below, in that case the revenue volumes (and related algebraic expressions) may be similarly defined, except that the roles of p and n are reversed. However, to avoid these restrictions on the demand distribution we initially focus attention on the all-or-nothing distribution defined in Section I. Here, for a given output, q_0, revenue is defined by:

$$R(q_0) = q_0 \cdot C(n,q_0) \cdot n, \tag{7}$$

where C is the (per-unit) all-or-nothing charge for all units in that output bundle. The optimal value of n and associated charge C may

be obtained directly from the first-order condition:

$$\frac{\partial R(q_0)/\text{OAN}}{\partial n} = q_0\left[n \cdot \frac{\partial C}{\partial n} + C(n,q_0)\right] = 0. \tag{8}$$

This must be true for all values of q and defines n as a function of q for optimal values of C. It is straightforward to proceed from here and obtain:

$$\text{MR}(q)/\text{OAN} = n \cdot C(n,q) + \frac{\partial R(q)}{\partial n} dn^*, \tag{9}$$

where dn^* is the change in n when q is increased by one unit with C held constant (at its previous optimal level).[12]

Operationally, marginal revenue thus defined may be obtained in just two steps: First, locate the optimal per-unit charge for a given output level (which gives us $n \cdot C[n,q]$), and second, note the change in the number of consumers when output is increased by one unit and the charge (per-unit) is held constant (which yields the second part of eq. [9]). Evidently, *MR(q)/OAN is not much more difficult to determine than MR(q)/OUP.*

Of more immediate interest, however, is the revenue yield of all-or-nothing charges relative to OUP. Insight may be obtained by limiting attention to the case where individual demand curves do not intersect. In these circumstances, $R(q_0)/\text{OUP}$ involves choosing p to maximize $\int_0^{q_0} p \cdot n(p,q)dq$ while $R(q_0)/\text{OAN}$ involves choosing n to maximize $\int_0^{q_0} n \cdot p(n,q)dq$ (in the nonintersecting case $p[n,q]$ being, for given n, the nth individual's inverse demand function). The natural symmetry between the two procedures leads directly to two observations. First, for any distribution which is symmetrical in p and n (i.e., so that if $p = a$, $n = b$ satisfies $n = n[p,q_0]$, so must $p = b$ and $n = a$) for some scale of measurement of these variables, the revenues must be identical; second, the existence of some distribution for which $R(q)/\text{OAN} > R(q)/\text{OUP}$ implies the existence of a distribution yielding a contrary result. Thus, *OAN may be superior or inferior to OUP in revenue terms depending upon the characteristics of the demand distribution.* Moreover, the general properties of the two strategies may be quite similar. For example, by arguments directly analogous to those used in propositions 1 and 2 it follows that *the optimal value of C may increase as output increases, but not if the all-or-nothing distribution is both regular and weakly restricted.* It is also possible that the number consuming a given bundle, $n(q)/\text{OAN}$, may sometimes increase as output increases.[13]

[12] See again Courant (1960). As an approximation, the linear form MR(q)/OAN = MC/2 may be useful in estimating whether the unit output expansion should be tried.

[13] This can be seen easily for a distribution with nonintersecting demands (but no other restrictions), since then OAN analysis can be conducted with the basic demand

Given that selection of the Optimal All-or-Nothing Charge may be only slightly more difficult than selection of the Optimal Uniform Price, there is virtually no theoretical basis for choosing between the two procedures. This should cause little surprise because the degree of complexity underlying the maximization problem faced by the entrepreneur is identical in both cases. Alternatively, insight into the ambiguity in revenue ranking may be obtained from the realization that *$C(q)$OAN may be $\lesseqgtr p(q)$/OUP depending upon the distribution* and (compared with a uniform per-unit price equal to C) that *OAN involves the additional consumption of (lower-valued) units by some individuals while excluding all (higher-valued) consumption by others.*[14]

Two-Part Pricing

The characteristic feature of Two-Part Pricing (TPP) is the combined use of an entrance (or license) fee and a uniform per-unit price—a combination which is commonly used in the marketing of recreational and entertainment services. The essence of the TPP case can be captured in the following:

PROPOSITION 3: The introduction of a license fee in conjunction with uniform per-unit pricing may reduce revenue. However, if the demand distribution is both regular and weakly restricted and $p(q)$/OUP is used as the per-unit price, revenue is necessarily increased.

PROOF: With uniform per-unit pricing, more individuals consume the "initial" unit of output than any other unit. The marginal introduction of a license fee may be viewed as raising the price of the initial output unit without affecting the price or consumption of later units. The revenue effect will, therefore, be

$$\frac{\partial[p(n,0) \cdot n(p,0)]}{\partial p},$$

and even for a regular distribution this may be positive or negative depending upon whether the original per-unit price was below or above $p^*(0)$ (the price which maximizes revenue from the first unit alone). It follows directly from proposition 2, however, that if OUP is

---

distribution but with p and n reversed. Whatever holds for OUP, an analogous result holds for OAN. This similarity between OAN and OUP, moreover, suggests an analogue to MUP which might be termed "simple all-or-nothing pricing" (SAN). For output q, it involves setting a charge so that all who have positive marginal evaluations at that output are permitted to consume, the marginal consumer paying his total evaluation. $R(q)$/SAN $= R(q)$/MUP for nonintersecting demands with the distribution symmetrical in p and n.

[14] Indeed, for the linear distribution used in Section IV we have both that $C(q)$/OAN $= p(q)$/OUP and that the additional and excluded consumptions are identical. Therefore, $R(q)$/OAN $= R(q)$/OUP.

used and the distribution is both regular and weakly restricted, the revenue effect must be positive. □

That this proposition has two distinct components is quite intentional, since together they shed light on a trade-off which gives rise to the existence of an interesting possibility. From the first part of the proposition, if a given output level has a uniform price greater than $p^*(0)$, the introduction of an entrance fee will cause a greater loss of revenue from now-excluded consumers than is gained from entrance payments by the remaining consumers. From the second part of the proposition, the two effects yield a net revenue gain for a given output level for *any* uniform price *lower* than $p^*(0)$. Insight into the distinction between the two cases may be aided by regarding the first as involving a relatively high per-unit price and small number of consumers and the second as involving the reverse.[15] However, in lowering the per-unit price to make an entrance fee profitable (in an absolute sense) the departure from the uniform per-unit price $p(q)$/OUP may be sufficiently large to involve a greater loss of revenue than can ever be captured by the fee. Thus, *for some distributions it may never be profitable (relative to OUP) to use an entrance fee in conjunction with uniform per-unit pricing (i.e., TPP).*

Notwithstanding this interesting possibility, proposition 3 does suggest that under plausible conditions some form of two-part pricing will be superior to OUP, and our conceptual experiment of introducing a *marginal* entrance fee suggests there is no great difficulty in gaining *some* extra revenue. Indeed, if for any given output level the revenue-maximizing license fee and per-unit price could be determined, there would be little difficulty in locating the profit-maximizing output level, since we must have:

$$\mathrm{MR}(q)/\mathrm{TPP} = p \cdot n(p,q), \tag{10}$$

where p is the relevant per-unit price for output level q. This unsurprising result follows from arguments similar to those used to derive MR(q)/OUP above and, indeed, it is an interesting property of many marketing strategies that the marginal revenue function is defined by the revenue obtained solely in respect of the final unit of output. The reason is straightforward. The optimal setting of the parameters for any such strategy implies that marginal changes in those parameters will not change the revenue associated with any given output level. If a marginal increase in output results in only an incremental change in the price parameter applying to final production units, then revenue with respect to the original output level will be unchanged, and *any*

[15] Strictly speaking, allowance should also be made for the way n changes as p changes in the two situations.

change in overall revenue must be simply that amount obtained solely from the final output unit.[16]

Other results (related to those for OUP) follow almost immediately from (10). Except where marginal costs are zero, *TPP will involve rationing of some individuals by output; and the per-unit price component may increase with output* but it will not do so if the demand distribution is both regular and weakly restricted. In these latter circumstances it can also be established that the per-unit price element of TPP is less than $p(q)$/OUP.[17] However, even allowing for this last insight and despite the simplicity of the marginal revenue function, the entrepreneur would still need to undertake innumerable experimental variations in both the license fee and per-unit price components at each output level. *It remains unclear, therefore, whether TPP dominates OUP (and, a fortiori, OAN) in terms of overall (operational) profitability even when it does so in terms of potential revenue-raising capacity.*

Multipart Pricing

We now suppose that the producer considers confronting consumers with a schedule of prices in which different units (or bundles of units) attract a different price (though the same price schedule is applied to all consumers). This strategy might be exemplified (though perhaps somewhat imperfectly) by "quantity discount" offers for joint goods, such as are applied to a series of orchestral concerts, operas, or plays.

Obviously, the Optimal Multipart Pricing procedure (MPP) will involve pricing each unit separately at the price, $p^*(q)$, which maximizes revenue obtained solely in respect of that unit (although, as we observe later, for operational reasons the producer may offer different sized *bundles* of units with a composite charge, or average per-unit equivalent, equal to the sum of the revenue-maximizing separate prices for the units it contains). The price-output relationship is, therefore, immediately defined and, as with other strategies, *price may increase with output over some ranges.* However, in all cases the overall *trend* must involve prices decreasing with output, and for a weakly restricted distribution (this restriction directly relating to $p^*[q]$) the price-output schedule would necessarily be negatively sloped, with higher prices charged for the most intensively used ("lower level") units of output.

[16] The reader may wish to confirm why such arguments do not apply to the MR(q)/OAN function defined with respect to parameters of the all-or-nothing distribution.

[17] Given $\partial C/\partial q, \partial C/\partial n < 0$, the per-unit price component of TPP will be that which maximizes revenue in respect of the last $(q - q^*)$ units, q^* being the minimum consumption of any (nonexcluded) individual. From propositions 1 and 2 this price must be less than $p(q)$/OUP given the stated restrictions.

For the MPP strategy, marginal revenue is trivially defined as the maximum revenue obtainable solely in respect of the qth unit, $\{p^*(q) \cdot n[p^*(q), q]\}$, and, given the properties of the demand distribution in (1) above, MR(q)/MPP must decrease as output increases. Equally clearly, *$R(q)$/MPP must exceed $R(q)$/OUP, except in the special case where $p^*(q) = p^*(0)$ for all q, in which case the strategies are identical in all respects.*

This superiority of MPP might have been expected to extend straightforwardly to comparisons with other relatively simple strategies, such as OAN. Interestingly enough, no such exact result can be obtained unless the restriction is made that individual demand curves do not intersect. In these special circumstances, the marginal consumer of a bundle of units under OAN necessarily has the lowest marginal valuation of each and every unit in the bundle. The revenue obtained under OAN would, therefore, be equal to that obtained from a form of multipart pricing which makes the units available at the separate prices, $p_m(q)$, given by this marginal individual's demand curve. However, *except where (quite fortuitously) those prices are also the optimal $p^*(q)$ this OAN revenue must be less than that obtained under MPP.* In more general circumstances where demand curves do intersect, it can no longer be presumed that all individuals who would consume the entire bundle at the OAN composite charge would still do so if the units were priced separately at $p_m(q)$. Even though it is also true that some individuals who would not purchase the bundle at the composite charge would undertake *some* consumption at these separate prices, it remains possible that $R(q)$/OAN may exceed the revenue obtained from the multipart pricing arrangement using $p_m(q)$. Since we do not know by how much $R(q)$/MPP exceeds this latter arrangement, we cannot be certain that it in fact exceeds $R(q)$/OAN. And this uncertainty extends to comparisons between MPP and any strategy involving a lump-sum component, including TPP. It appears that *the general expectation that MPP will be the revenue-dominant strategy, while intuitively plausible, can only be established in a probabilistic sense.* Nonetheless (and notwithstanding some operational difficulties referred to below) MPP might be regarded as closest to a theoretical "best" procedure in revenue terms when interpersonal discrimination is ruled out.

Two further aspects of the MPP strategy might be noted. First, the notion of (a schedule of) prices related to "intensity of use" immediately suggests a comparison between the MPP solution and Oakland's (1974) competitive solution. Although for Oakland's case the price schedule is unambiguously positively related to output while our monopoly counterpart may characteristically involve a negative relationship between price and output, the similarities are more than merely superficial. For a given demand distribution and marginal cost schedule, the Oakland competitive output and the monopoly output

under MPP must be identical, since both solutions must be characterized by marginal cost being equated to the "total price" ($n \cdot p$) which maximizes revenue with respect to the last unit of output and, in both cases, this will in general be associated with rationing by output.[18] (Of course for Oakland no above-normal profit exists on any unit, and consumption by inframarginal consumers is greater for his solution than under MPP because inframarginal units bear a lower price in the competitive case.)

Second, however, given that the characteristic MPP solution involves a negatively sloped price schedule, the operational status of MPP differs markedly from that of Oakland's solution. For MPP we must assume that producers can acquire information on the relevant revenue-maximizing prices, can prevent resale (or exchange) of units, and can prevent consumers from "reversing intensity" (i.e., purchasing lower-priced units without first purchasing the higher-priced inframarginal units). Resale possibilities may be of little concern in the joint good context given the service nature of most of the examples. However, preventing reversals of intensity requires either that the producer can completely meter individual purchases or that he can present consumers with a choice between bundles of units (containing 1, 2, 3, and so on, up to q units) each with a composite charge equal to the sum of the revenue-maximizing prices for the separate units it contains (a strategy we might call MPP*). However, for MPP* to be made operational, producers must be able to determine the optimal set of composite charges (and hence the relevant $p^*[q]$), and this may be a great deal more complex than isolating the $p^*(q)$ when intensity reversals do not occur. In general, firms would initially have to reduce output to zero and proceed sequentially—first determining the revenue-maximizing charge, $p^*(1)$, for one unit of output, then producing a second unit and determining the optimal composite charge for a two-unit bundle (i.e., $CC^*[2] = p^*[1] + p^*[2]$), and so on.[19] If resale possibilities or problems of overcoming intensity reversals

[18] Oakland's output is the maximal quantity, q, for which $k \cdot p_k(q) = MC$ (ranking individuals $k = 1, \ldots, K$ so that $p_1[q]$ is the highest inverse demand). If $k \cdot p_k(q) > MC$ (for some $k > 1$) when q is such that $p_1(q) = MC$, output is expanded and so on. This is equivalent to increasing output whenever the maximum revenue from the marginal production unit exceeds MC and, unless q is maximal where $p_1(q) = MC$, at least the highest demand individual is rationed by output. A referee suggested that the equality of output for MPP and competition is analogous to the private good result for perfect discrimination and perfect competition. This may be potentially misleading: MPP is not perfect discrimination, and in general the MPP (and hence competitive) output is not efficient.

[19] Though this may be a protracted experiment and fairly costly in forgone-revenue terms (particularly if fixed costs are high), if the revenue-maximizing charge declines, firms may be able to minimize their losses by allowing consumption of further units at the per-unit price $p^*(1)$ while experimenting to find the relevant charges for bundles of two or more units. Similar but more complex procedures are required if the optimal charge may increase over some output ranges.

create insurmountable problems, then the producer will be obliged to set prices such that $p(q) \leq p(q + dq)$ for all q, and the most that can be said is that for a weakly restricted distribution, the optimal separate prices are in fact uniform and therefore coincident with OUP.

Summary of Results

The more complex, but uniformly applied, pricing strategies considered here retain two basic characteristics of optimal uniform per-unit pricing: All generally involve rationing by output, and all may generate a positive price-output relationship over some output ranges. For all the strategies, moreover, the relevant marginal revenue functions can be defined with little or no more difficulty than under OUP. Optimal All-or-Nothing pricing (OAN) is operationally little more complex than OUP, but there are no strong theoretical grounds for choosing between OAN and OUP in revenue terms. Selection of the Optimal Two-Part Price (TPP) is more difficult than selection of the OUP (and OAN), and its revenue dominance is assured only for restricted (though plausible) forms of the demand distribution. Optimal Multipart Pricing (MPP) is revenue dominant over OUP, but its dominance over OAN and TPP cannot be universally guaranteed, and its position as the "theoretical best" strategy can only be established in probabilistic terms. The MPP output level is, for given cost and demand conditions, identical to Oakland's competitive output, but for operational reasons MPP may be of somewhat limited commercial applicability.

IV. Illustrative Analysis with a Specific Distributional Form

Both to illustrate the methodology used in the general analysis and to consider some of the welfare properties of monopoly provision, we now analyze the results that emerge when a specific demand distribution is used. In choosing a distributional form, account has been taken of three factors. First, there would be heuristic gains from adopting a framework which could perform an expositional role similar to that of linear demand curves in traditional demand analysis. Second, the chosen distributional form should exhibit the more fundamental characteristics of distributions likely to be met in practice. Third, the framework should not be devoid of operational content. Consideration of these factors leads more or less unequivocally to the choice of a distribution similar to that shown in figure 2 but for which the surface is linear.[20] To enable the framework to include all-or-

[20] On this, and for a further application of results obtainable from a linear demand distribution, see Burns (1979).

nothing charges we also assume that demand curves do not intersect and that income effects may be regarded as negligible.

Within this framework, the bounding values P, Q, and N completely define the distribution and may be treated as (empirically determinable) constants throughout the analysis which follows; consideration of alternative pricing procedures can be simplified considerably by drawing on the proportionality properties of similar triangles. In this respect it is convenient to define

$$\hat{p} = \frac{p}{P}, \hat{q} = \frac{q}{Q}, \text{ and } \hat{n} = \frac{n}{N}, \tag{11}$$

where p, q, and n represent specific values of price, output, and numbers of individuals. The actual derivation of the required revenue functions and pricing rules for a range of pricing strategies has been presented in some detail in Burns (1979), and the summary information contained here in table 1 and figure 4 is sufficient for present purposes.

In table 1 seven alternative strategies are defined, the associated revenue functions and pricing rules being shown for each of these strategies. The revenue functions have been presented in a form that makes self-evident the question of relative profitability (for given output levels) and illustrates certain results obtained earlier, such as the general dominance of OUP and OAN over MUP and SAN. These functions also provide a neat illustration of two further points. First, and unsurprisingly, revenue (for given output) is positively correlated with the degree of complexity of the strategy adopted. Second, broadly speaking, schemes of comparable complexity should be expected to yield similar revenues: for example, we can see that there is nothing to choose between OUP and OAN, although nonlinear distributions could be constructed (but probably not empirically determined) to favor either of these procedures. It is also apparent from the revenue functions that the returns to increased complexity of strategy diminish very rapidly, particularly at lower output levels.

Figure 4 is fundamental in illustrating the basic methodology suggested in this paper. Entrepreneurial decision making concerning output and price determination relies heavily on information about the marginal revenue function of a chosen strategy and, of course, on the relationship of such a function to (marginal) cost behavior: figure 4 contains the derived marginal revenue functions for each strategy. Also shown is a schedule indicating the sum of marginal valuations at different output levels: the intersection of this schedule with a marginal cost function would determine the optimally efficient output. Assuming, initially, constant marginal costs, the results for a linear distribution offer a clear warning that, over a substantial range of

TABLE 1

ALTERNATIVE PRICING STRATEGIES: DEFINITIONS AND REVENUE FUNCTIONS FOR A LINEAR DISTRIBUTION

Definition	Total Revenue	Marginal Revenue	Pricing Rule
Maximum Uniform Pricing—MUP: charging the maximum uniform price (uniform across output and individuals) such that some individual will wish to consume the entire output	$\hat{q}(2-\hat{q})^2PQN/16 - \hat{q}(2-3\hat{q})^2PQN/16$	$\hat{q}(2-3\hat{q})PN/2$	$p = (1-\hat{q})P$
Simple All-or-Nothing Charging—SAN: selecting the all-or-nothing charge for a bundle of q units such that all individuals who have a nonnegative evaluation of the qth unit shall be willing to pay that charge	...	...	$C = \hat{q}P/2$
Optimal Uniform Pricing—OUP: for a given output level q, charging the uniform price that maximizes revenue	$\hat{q}(2-\hat{q})^2PQN/16$	$(\hat{q}-1)^2PN/4 - \hat{q}^2PN/16$	$p = (2-\hat{q})P/4$
Optimal All-or-Nothing Charging—OAN: for a given output level q, selecting the all-or-nothing charge that maximizes revenue	...	...	$C = (2-\hat{q})P/4$
Optimal Two-Part Pricing—TPP: for a given output choosing the combination of lump-sum charge plus a uniform price that maximizes revenue	$\hat{q}(2-\hat{q})^2PQN/16 + \hat{q}^3PQN/64$	$(\hat{q}-1)^2PN/4 - \hat{q}^2PN/64$	Entrance fee $\hat{q}P/4$ $p = (4-3\hat{q})P/8$
Separate Pricing of Bundles—SPB: jointly choosing the optimal partition of output into two bundles and an optimal all-or-nothing charge for each bundle	.	...	Optimal bundles are equal size: $p_1 = (4-\hat{q})P/8$ $p_2 = (4-3\hat{q})P/8$
Optimal Multipart Pricing—MPP (or MPP*): assuming that each unit of output (or each consumption level—MPP*) may be priced (charged) separately, selecting those prices (charges) for each unit (bundle) that maximize revenue	$\hat{q}(2-\hat{q})^2PQN/16 + \hat{q}^3PQN/48$	$(\hat{q}-1)^2PN/4$	$p_i = (1-\hat{q}_i)P/2$

NOTE.—Both total and marginal revenue functions are defined for $\hat{q} = q/Q$. The algebraic expressions are based upon those derived in Burns (1979) and have been presented in comparable rather than simplified forms.

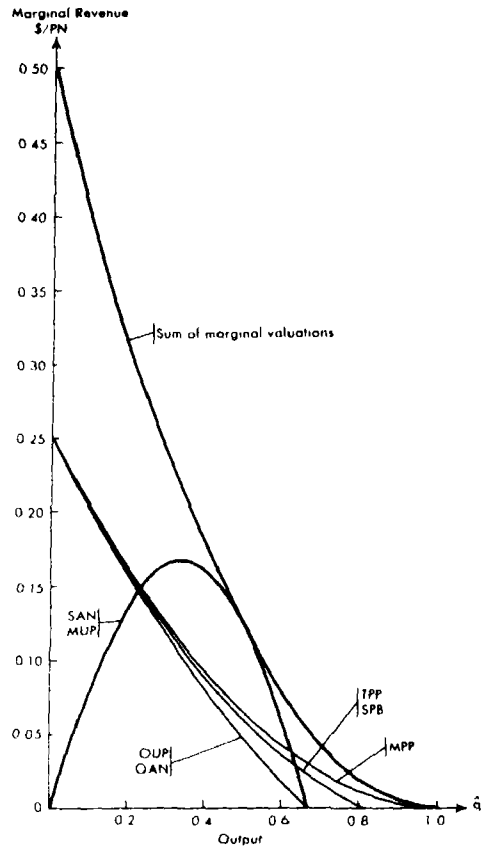


FIG. 4 — Marginal revenue functions for a linear distribution

(higher) marginal cost values, there may be negligible differences in the outputs achieved by different procedures of varying complexity and that the outputs attained by even the most complex procedures may be considerably less than the optimally efficient output. On the other hand, at lower levels of marginal cost the difference in output levels becomes more pronounced, indicative of the fact that more complex procedures achieve their maximum possible revenues at higher output levels, but in general all procedures more nearly approach the optimally efficient output level, especially the more sophisticated ones. Finally, noting that the monotonicity of MR(q)/MPP is guaranteed while that of other marginal revenue functions is plausible,[21] a situation of decreasing marginal costs will increase the

[21] E.g., although MR(q)/OUP may sometimes increase with output, its general trend must be downward (see Section II, "Some General Properties of Revenue Functions

divergence in output levels, while increasing marginal costs would have the converse effect.

In conclusion, we offer a cautionary remark about comparisons between privately achieved and optimally efficient output levels. Our results indicate (not surprisingly) that suboptimal output levels can be expected for all of the private procedures considered. While these results are relevant to the concern shown by a number of authors who have made such comparisons in other contexts (and have sometimes claimed that efficient outputs might be secured privately), it is clear that comparisons of output levels alone are of little interest. In the case of the linear distribution, outputs under OUP and OAN would be identical, but the efficiency implications of the two procedures are significantly different. Indeed, one of the insights generated by the approach advanced in this paper is that the replacement of a uniform per-unit pricing strategy with a (comparably complex) all-or-nothing arrangement will succeed only in substituting the consumption of lower-valued units of output by higher-demand individuals for the consumption of higher-valued units of output by lower-demand individuals. In general, efficiency comparisons must be based on the degree of consumption rationing (whether by price or by output) that occurs for all consumers vis-à-vis the completely efficient outcome in which all individuals would consume all units produced and output would be such that the sum of marginal valuations equals marginal cost.

V. Concluding Remarks

A major contribution of this paper is its demonstration that the complete demand distribution provides a novel and powerful basis for analyzing private provision of price-excludable public (or joint) goods. For a monopoly producer of such goods we have identified, under surprisingly general conditions, the characteristics of the revenue functions associated with various pricing strategies, indicated the likely relative profitability of those strategies, and prescribed procedures which might aid firms in isolating profit-maximizing price and output for any chosen strategy. Moreover, the specific example we provide, based on a linear demand distribution, not only serves to illustrate our general results, but also in practice may constitute an important empirical basis for generating first approximations for output and price levels.

---

under OUP"). Moreover, given MPP as the theoretical best strategy, MR(q)/MPP would be the boundary form toward which other MR functions tend (as complexity of the pricing strategy is increased), and MR(q)/MPP is unambiguously negatively sloped.

Using this linear demand distribution, we have also sketched out some efficiency implications of our results, and, in this regard, another important application of our general approach is highlighted. For joint goods, overall efficiency depends on how output is "rationed" among consumers as much as on the level of output per se. A framework containing information on the distribution of demands clearly is indispensable to a comprehensive analysis of the efficiency of private (or, indeed, public) provision of joint goods. In fact, our framework could be used profitably as a basis for a general analysis of joint goods provision under competitive conditions and other market or nonmarket structures. Moreover, the demand distribution approach is equally essential to analysis of "discriminatory" pricing strategies for private goods, the necessary extension of the analysis presented here being particularly straightforward when constant marginal costs prevail.

References

Auster, Richard D. "Private Markets in Public Goods (or Qualities)." *Q.J.E.* 91 (August 1977): 419–30.

Berglund, Sture. "Toward a Formal Discussion of Quantity Discounts: A Dualistic Approach to Price Differentiation." *Swedish J. Econ.* 66 (December 1964): 274–85.

Brennan, Geoffrey, and Walsh, Cliff. "A Monopoly Model of Public Goods Provision. The Uniform Pricing Case." Seminar Paper no. 70, Monash Univ., Dept. Econ., January 1978

———. "Market Provision of Public Goods: A Monopoly Version of the Oakland Model." *Finanzarchiv* 37, no. 3 (1979): 385–95.

Buchanan, James M. "The Theory of Monopolistic Quantity Discounts." *Rev. Econ. Studies* 20, no. 3 (1952–53): 199–208.

———. "Public Goods in Theory and Practice: A Note on the Minasian-Samuelson Discussion." *J. Law and Econ.* 10 (October 1967): 193–97.

Burns, Michael E. "Discrimination and Efficiency in the Pricing of Public Goods." Seminar Paper no. 82, Monash Univ., Dept. Econ., 1979.

Courant, Richard. *Differential and Integral Calculus.* Vol. 2. New York: Wiley, 1960.

Demsetz, Harold. "The Private Production of Public Goods." *J. Law and Econ.* 13 (October 1970): 293–306.

Dupuit, Jules. "On the Measure of the Utility of Public Works." Translated in *Internat. Econ. Papers*, no. 2 (1952), pp. 83–110 (*De l'utilitie et de sa mesure*, 1844; reprint ed. [with comments by M. de Bernardi and L. Einaudi], Turin, 1934).

Gabor, André. "A Note on Block Tariffs." *Rev. Econ. Studies* 23, no. 1 (1955–56): 32–41.

Lee, Dwight R. "Discrimination and Efficiency in the Pricing of Public Goods." *J. Law and Econ.* 20 (October 1977): 403–20.

Ng, Yew-Kwang, and Weisser, Mendel. "Optimal Pricing with a Budget Constraint: The Case of the Two-Part Tariff." *Rev. Econ. Studies* 41 (July 1974): 337–45.

Oakland, William H. "Public Goods, Perfect Competition, and Underproduction." *J.P.E.* 82, no. 5 (September/October 1974): 927–39.

Puu, Tõnu. "A Note about Second-Order Conditions for the Monopolist's Optimum at Differentiated Prices." *Swedish J. Econ.* 66 (September 1964): 208–14.

Samuelson, Paul A. "The Pure Theory of Public Expenditure." *Rev. Econ. and Statis.* 36 (November 1954): 387–89.

———. "Diagrammatic Exposition of a Theory of Public Expenditure." *Rev. Econ. and Statis.* 37 (November 1955): 350–56.

Thompson, Earl A. "The Perfectly Competitive Production of Collective Goods." *Rev. Econ. and Statis.* 50 (February 1968): 1–12.

Yamey, Basil. "Monopolistic Price Discrimination and Economic Welfare." *J. Law and Econ.* 17 (October 1974): 377–80.

Comments

Recalculating the Scientific Tariff

Bruce R. Bolnick

Harvard Institute for International Development

In part 2 of his classic 1960 paper, "The Cost of Protection and the Scientific Tariff," Harry G. Johnson analyzed the second-best[1] problem of using tariffs to promote "noneconomic" objectives commonly suggested as arguments for protection. Very simply, Johnson showed how to compute the balance between marginal benefits and marginal costs using Harberger-type welfare triangles. He illustrated the methodology for five selected noneconomic objectives: (*a*) a tariff to promote self-sufficiency and independence; (*b*) a tariff to promote diversification, industrialization, or agriculturalization; (*c*) a tariff to promote a "way of life"; (*d*) a tariff to increase military preparedness; and (*e*) a bargaining tariff.

Johnson assumed that world prices of importables are given, that domestic production of importables is positive before and after imposition of the tariff, and that cross-elasticity effects can be neglected.[2] Using linear demand and supply curves and free-trade values for market parameters, Johnson computed the marginal cost of an ad valorem tariff (t) to be precisely

$$MC = tP\epsilon + tC\eta, \tag{1}$$

where P and C are the free-trade levels of domestic production and consumption, and ϵ and η are the compensated own-price elasticities

I would like to express my great appreciation to Professor Edward Tower for invaluable discussions and suggestions.

[1] Strictly speaking, the use of tariffs for such purposes is not necessarily *second* best. It may be third best if the "noneconomic" objective implies a divergence in factor markets; or it may be first best if a trade divergence is involved (see Corden 1974, chap. 2).

[2] Units of output are assumed to be defined so that the world price is unity. This assumption, along with the others listed in the text, serves to simplify computations.

[*Journal of Political Economy*, 1981, vol. 89, no. 1]
© 1981 by The University of Chicago. 0022-3808/81/8901-0011$01.50

of supply and demand. (I have dropped Johnson's subscript i, indicating parameters for good i, in order to simplify the notation.)

Given any specific tariff-related policy objective, the marginal benefits (MB) can be computed as a function of t and the parameters P, C, ϵ, and η. Johnson's methodology then involves setting the ratio MC/MB (call this Z) equal to unity and inferring the characteristics of the scientific tariff structure from the resulting expression.[3] However, the condition $Z = 1$ is neither necessary nor sufficient for determining the scientific tariff unless the prohibitive tariff (which satisfies $Z = 1$ only by accident) can somehow be ruled out of consideration as the second-best optimum. Furthermore, even excluding the prohibitive tariff from consideration, $Z = 1$ is sufficient only when it has a unique root in t, which will be true only for a special class of policy objectives (which define MB). Thus the scientific tariff can be inferred from the first-order conditions in only a limited set of circumstances—even if one adheres to the simplifying assumptions underlying Johnson's analysis.

These problems can be illustrated using Johnson's own case of "self-sufficiency," which he chose to measure by "the proportion of consumption supplied from domestic production." This gives

$$MB = \beta \frac{P}{C} (\epsilon + \eta)(1 - t\eta)^{-2}, \tag{2}$$

where β is the policy weight attached to marginal units of the objective, in numeraire units.[4] Thus

$$Z = \frac{MC}{MB} = \frac{tC}{\beta} \left(1 + \frac{C - P}{P} \frac{\eta}{\epsilon + \eta}\right)(1 - t\eta)^2, \tag{3}$$

which clearly is cubic in t and can therefore generate multiple roots[5] for the equation $Z = 1$. Note too that the prohibitive tariff (for which $C = P$) is at the level t_p satisfying $t_p = (C - P)/(C\eta + P\epsilon)$.

Figure 1 shows how MB and MC vary with t. Here $MB(\beta_1)$ is the marginal benefit line for a low value of β, in which case the scientific tariff (t_{s1} where $Z = 1$) is unambiguously preferred to the prohibitive tariff (t_p). For some policy weight β_2, $MB(\beta_2)$ will satisfy the property

[3] Specifically, from the total differential of Z one can determine the dt necessary to offset dX ($X = \epsilon, \eta, P$, or C), thereby characterizing the differences in t as the market parameters vary across goods or over time.

[4] I have introduced the parameter β into Johnson's model to make MB commensurate with MC, in order to clarify the discussion. This parameter is assumed to be constant, purely as a simplifying device.

[5] Johnson, in his (1960) article, avoided confronting this problem by using a linear approximation to (3) in his analysis. This device is permissible if and only if one can presume that the scientific tariff will be small, but that presumption cannot generally be valid since the whole point of the exercise is to *find* the level for the scientific tariff.

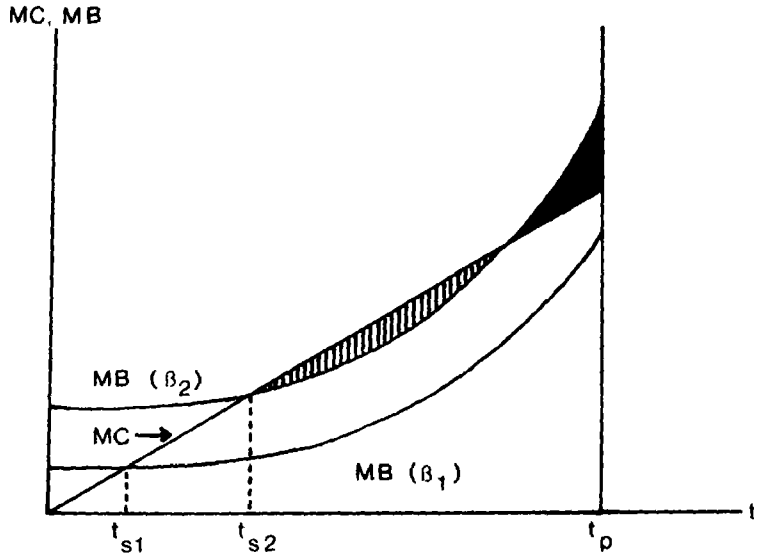


FIG. 1

that the solid area equals the slashed area, so the policymaker is indifferent between a tariff of t_{s2} (where $Z = 1$ given C, P, ϵ, and η) and one which is prohibitive ($\geq t_p$). For any $\beta > \beta_2$ a prohibitive tariff will always be preferred. Thus a small increase in β may cause a finite jump in the scientific tariff.[6]

Similarly, a small change in any of the market parameters may cause a finite jump in the scientific tariff (by altering MB, MC, and t_p). For example, beginning with MC and $MB(\beta_2)$ as shown in figure 1, it is clear from equation (3) that an increase in ϵ will shift MB up relative to MC, causing the scientific tariff to jump from t_{s2} to t_p (which itself is reduced in magnitude since $dt_p/d\epsilon < 0$). The possibility of such discontinuities is completely absent from Johnson's well-behaved but inappropriate formula. Furthermore, any further increase in ϵ will reduce the scientific (i.e., prohibitive) tariff, contrary to Johnson's conclusion that "the implied scientific tariff structure entails tariff rates which are higher the higher the elasticity of supply . . ." (p. 343).

In short, Johnson's methodology applies only to a limited set of circumstances; more generally, the first-order conditions are neither necessary nor sufficient for identifying the "scientific tariff."

[6] For example, if $\beta = 1$, $C = 10$, $P = 1$, and $\eta = \epsilon = 1$, then $t_p = .82$ while $t_{s2} < .33$ (since t_{s2} must be to the left of the peak of the cubic Z function, which occurs at $1/3\eta$). Thus a slight change in the planner's preference for protection, or any of the market parameters, could conceivably cause the scientific tariff to leap from below 33 percent up to 82 percent.

Moreover, the structure of protection implied by the correct calculation may differ dramatically from that implied by Johnson's procedure.

References

Corden, W. M. *Trade Policy and Economic Welfare*. Oxford: Clarendon, 1974.
Johnson, Harry G. "The Cost of Protection and the Scientific Tariff." *J.P.E.* 68, no. 4 (August 1960): 327–45.

Discount Rate and Wealth

G. S. Laumas

Illinois State University

In a recent issue of this *Journal*, Mohabbat and Simos (1978) provided estimates of the rate of discount using Kendrick's (1976) series of total private wealth. Since Kendrick also provides estimates of the division of total wealth between human, W_h, and nonhuman, W_n, wealth, it should be of interest to investigate how the discount rates differ between the two components of wealth.

In order to accomplish this, it is necessary to divide the total income into two parts—one corresponding to the human wealth, Y_h, and the other to the nonhuman wealth, Y_n. Kendrick (1976, p. 22) and Christensen and Jorgenson (1973) provide an outline of the method for such a division of the total income. Following them, the time series for Y_h and Y_n were estimated. The estimated series closely approximate the estimates for "labor compensation" and "property compensation" given by Kendrick and Christensen and Jorgenson.

As suggested in Mohabbat and Simos, the time-series estimates of the rate of discount for W_h and W_n were computed using the varying parameter regression technique. These series are given in table 1.

The obvious inference that can be drawn from the results given in table 1 is that the discount rate on W_h is consistently higher than the rate on W_n. This evidence is important by itself. In the absence of direct empirical estimates, as provided here, economists had to make assumptions about the behavior of the two rates (see Pesek and Saving 1967).

Additionally, since the major portion of W_h is education, the results provide further evidence on what Welch has called "one of the most important phenomena of our time . . . that the rates of return to investments in schooling have failed to decline under the pressure of rapidly rising average educational levels" (Welch 1970, p. 54). The existence of a relatively higher rate on W_h confirms the conclusions about the behavior of the rate of return on education arrived at by Griliches (1970, 1977), Welch (1970), Becker (1975), and others.

[*Journal of Political Economy* 1981, vol. 89, no. 1]
© 1981 by The University of Chicago 0022-3808/81/8901-0010$01.50

TABLE I

TIME-SERIES ESTIMATES OF THE DISCOUNT RATE ON HUMAN AND NONHUMAN WEALTH, UNITED STATES, 1929 TO 1969

Year	Discount Rate on Human Wealth	Discount Rate on Nonhuman Wealth
1969	.15	.08
1968	.15	.08
1967	.15	.09
1966	.15	.09
1965	.15	09
1964	15	.09
1963	.15	.08
1962	16	08
1961	16	08
1960	.16	08
1959	16	.08
1958	.16	08
1957	17	08
1956	17	08
1955	17	.08
1954	17	08
1953	17	08
1952	.17	08
1951	.17	08
1950	17	08
1949	17	08
1948	17	08
1947	17	08
1946	18	08
1945	20	.08
1944	.21	08
1943	.21	08
1942	20	.08
1941	18	.07
1940	16	08
1939	16	.08
1938	.16	07
1937	16	.07
1936	16	08
1935	14	07
1934	.14	06
1933	13	.06
1932	.13	.05
1931	14	07
1930	.15	08
1929	16	.09

References

Becker, Gary S. *Human Capital* 2d ed. New York: Columbia Univ. Press (for Nat. Bur. Econ. Res.), 1975.

Christensen, Laurits R., and Jorgenson, Dale W. "U.S. Income, Saving, and Wealth, 1929–1969." *Rev. Income and Wealth* 19 (December 1973): 329–62.

Griliches, Zvi. "Notes on the Role of Education in Production Functions and Growth Accounting." In *Education and Income*, edited by W. Lee Hansen. Studies in Income and Wealth. Vol. 35. New York: Columbia Univ. Press (for Nat. Bur. Econ. Res.), 1970.

———. "Estimating the Returns to Schooling: Some Econometric Problems." *Econometrica* 45, no. 1 (January 1977): 1–22.

Kendrick, John W. *The Formation and Stocks of Total Capital*. New York: Columbia Univ. Press (for Nat. Bur. Econ. Res.), 1976.

Mohabbat, Khan A., and Simos, Evangelos E. "Consumer Horizon: Reconsidered." *J.P.E.* 86, no. 3 (June 1978): 539–41.

Pesek, Boris P., and Saving, Thomas R. *Money, Wealth, and Economic Theory*. New York: Macmillan, 1967.

Welch, Finis. "Education in Production." *J.P.E.* 78, no. 1 (January/February 1970). 35–59.

Book Reviews

Money, Inflation, and the Bank of Canada: An Analysis of Canadian Monetary Policy from 1970 to Early 1975. By THOMAS J. COURCHENE.
Montreal: C. D. Howe Research Institute, 1976. Pp. 290. $5.00.

For 12 years, from 1950 to 1962, while other industrialized countries were operating a fixed exchange-rate regime under the Bretton Woods par value system, Canada was a maverick floater that provided economists with a laboratory in which the practical problems of implementing monetary and fiscal policy under a regime of flexible exchange rates could be observed and analyzed. Prior to the advent of greater exchange-rate flexibility among the developed countries in the early 1970s, the Canadian experience was frequently cited as an example of the degree to which, through the judicious use of monetary policy, a flexible exchange rate could be kept stable with a minimum of official intervention in the foreign exchange market. It is no coincidence that two Canadians, Robert A. Mundell and Harry G. Johnson, were major contributors to the early literature on the appropriate use of monetary and fiscal policy under alternative exchange-rate regimes.

During the late 1950s and early 1960s, debate on the relative merits of fixed versus floating exchange rates was waged within Canada not as an abstract intellectual exercise, but as a policy issue of immediate practical interest to all Canadians. Canada returned to the fold of fixed exchange-rate countries in 1962 and remained there for 8 years. When she again chose to break from the par value system on June 1, 1970 she briefly appeared once more as the odd man out. By 1974, however, the accumulating pressures of widely divergent inflation rates among countries, continuing U.S. balance of payments deficits, and rapid increases in the prices of basic commodities and energy had led to the collapse of the par value system, leaving the larger industrial countries with little choice but to opt for the present system of "managed floating" exchange rates. In this context, the experience of the Bank of Canada in conducting monetary policy under alternative exchange-rate regimes is of considerable importance to students of monetary economics and policymakers alike. It is hardly surprising that, despite its relative youth (it was established in 1934), the Bank of Canada has been the subject not only of detailed research by academic economists, but also of lively and sometimes acrimonious public debate.

Thomas J. Courchene's careful study of the Bank of Canada during the crucial and fascinating period 1970–75 is a welcome addition to a substantial literature which includes works by Neufeld (1955), Johnson and Winder

Journal of Political Economy, 1981, vol. 89, no. 1)
© 1981 by The University of Chicago

(1962), Reuber (1964), the Royal Commission on Banking and Finance (1964), and, perhaps the best known, Paul Wonnacott's study *The Canadian Dollar, 1948–1962* (1965). Like these earlier works, Courchene's book attempts to analyze and explain the policies of the Bank of Canada within the context of standard concepts of macroeconomics and monetary theory. But Courchene's book is far from being a mere update of earlier research. His work differs in at least two ways. First, he seeks to explain his views not only to other economists but also to interested laymen, a desire which is the natural reflection of the position occupied by the Bank of Canada at center stage in the policymaking theater. Second, he approaches the analysis of monetary policy from an avowedly "monetarist" perspective. These considerations govern the structure of his work.

The book is divided into four main sections. Part 1 is devoted to an elementary elaboration of several important aspects of the monetarist theory of a small open economy under fixed and flexible exchange rates. Courchene emphasizes the distinction between nominal and real stocks of money and describes the different forces that act to equate money demand and money supply under alternative exchange-rate regimes. He then goes on to describe the concept of the natural rate of unemployment and spells out its implications for what monetary policy can and cannot do. Part 1 is completed by a brief description of alternative approaches to monetary theory, particularly the neo-Keynesian approach

Part 2 of the book treats the theory of central banking in general and the Bank of Canada's approach in particular. Courchene first considers the various indicators of monetary stringency or ease that were used in Canada during the 1970–75 period—nominal interest rates, the liquid asset ratio of commercial banks, M1 growth, etc. He argues that the Bank of Canada's choice of policy instruments during the period was a logical consequence of its theory of how the Canadian financial system operated and particularly of the role of market imperfections in allowing the authorities to influence both "the cost of credit" and "credit conditions" to some extent independently. Throughout, Courchene has chosen the technique of comparing actual Bank of Canada policies with what would have been ideal from the point of view of monetarist theory. Such an approach is, of course, consistent with his objective, but it does involve compromises. Monetary economists of a neo-Keynesian bent will not be entirely satisfied with Courchene's rather sparse treatment of the implications of the Keynesian model for monetary policy. Courchene rightly argues that his purpose is not to create a straw man out of the neo-Keynesian approach, but it is still left to the reader to decide whether policies that do not appear optimal when viewed from a strict monetarist perspective are more reasonable from the point of view of the model that the Canadian monetary authorities had in mind.

Fortunately, this question turns out to be somewhat beside the point, for the real meat of Courchene's book is to be found in parts 3 and 4, which present a very detailed description and analysis of Canadian monetary policy during 1970–75. After an initial chapter covering the whole period, Courchene devotes a chapter to developments during each year. This method of organization, while ideal for a reference work, makes for rather heavy going when the book is being read straight through because it results in excessive repetition. Nevertheless, Courchene's achievement in these sections is impressive. He manages to present a careful and incisive analysis that integrates a great deal of statistical material on the economy's performance with the

authorities' own public explanations of their policies, drawn mainly but not exclusively from Bank of Canada publications. As a result, the comparison of actual policy with the monetarist paradigm becomes less important than the detailed description of events.

Courchene's thesis is that the floating of the Canadian dollar in June 1970 was caused by the Bank of Canada's attempt to pursue a less inflationary policy than the rest of the world, a policy that could only succeed if the exchange rate was flexible. But Courchene argues that "when the rate floated and Canada was finally allowed the independence to pursue its own policies with respect to the behavior of prices, the concern over inflation was immediately jettisoned in favor of ensuring that the exchange rate was set at an 'appropriate' level" (p. 161). This change in emphasis, in Courchene's view, ushered in a period of excessive monetary expansion that persisted from 1971 until 1975. All this is supported by the marshaling of an impressive array of facts and figures. One may choose to disagree with Courchene's views about the monetary authorities' motives and intentions, but it is not easy to dispute the broad outlines of his argument. Along the way, the reader is treated to numerous interesting details: the difficulties experienced by the authorities in using indicators and instruments based on the asset side of commercial bank balance sheets; the subtle web of problems arising out of the federal government's cash deficit and the use of Canada Savings Bond issues to finance its needs; and reliance on the transfer of government deposits, rather than standard open-market operations, to alter the level of bank reserves.

On the basis of the evidence, few would dispute Courchene's conclusion that rapid rates of monetary expansion fueled inflation in Canada in 1971–75. But this criticism is somewhat tempered when the Bank of Canada's policies are compared with the actual track record of central banks in other countries during the same period rather than with the monetarist paradigm. Furthermore, by late 1975 the Bank of Canada, like several other central banks, had announced its intention to place greater emphasis on quantitative monetary targets, a fact recognized by Courchene in a postscript apparently inserted as his book went to press. Whatever the importance of these shifts in the thrust of monetary policy, Courchene's book will undoubtedly become a standard reference work for the 1970–75 period, and the thoroughness with which he has analyzed monetary policy in Canada will serve as a firm basis for the continuing debate on this subject.

MALCOLM KNIGHT

International Monetary Fund

References

Johnson, Harry G., and Winder, John W. L. *Lags in the Effects of Monetary Policy in Canada*. Working Paper. Ottawa: Queen's Printer (for Royal Commission Banking and Finance), 1962.

Neufeld, Edward P. *Bank of Canada Operations, 1935–54*. Toronto: Univ. Toronto Press, 1955.

Reuber, Grant L. "The Objectives of Canadian Monetary Policy, 1949–61: Empirical 'Trade-Offs' and the Reaction Function of the Authorities." *J.P.E.* 72, no. 2 (April 1964): 109–32.

Royal Commission on Banking and Finance. *Report.* Ottawa: Queen's Printer, 1964.

Wonnacott, Paul. *The Canadian Dollar, 1948–1962.* Toronto: Univ. Toronto Press, 1965.

A Life for Sound Money: Per Jacobsson. His Biography. By ERIN E. JACOBSSON. Oxford: Clarendon Press, 1979. Pp. xxv + 428. £16.

Erin Jacobsson has written a lively and readable life of her father, covering the many facets of a remarkable career. She did not lack for material: 6 meters of correspondence; a diary covering over 50 years in such detail that for the later years it would fill several volumes; some 500 articles; innumerable speeches and reports, including the greater part of the Bank for International Settlements (BIS) Annual Reports for 25 years. This is drawn upon to paint a picture of the man and his many activities over the years, along with a necessarily brief account of the background to some of those activities. A concluding chapter deals with his economic thought.

From the time he joined the League of Nations Secretariat in London in the spring of 1920, until his death in office as managing director of the IMF in 1963, Per Jacobsson was almost continuously in the service of a key international organization. There was a break of 2½ years in 1929–31 when he returned to Sweden. There for a time he was economic adviser to Ivar Kreuger, the Match King, not long before the latter's suicide and the sensational collapse of Kreuger and Toll. But even in those years he spent much of his time abroad on international business. Wherever he went as a wandering economic adviser he seemed to be in the thick of the action, with an extraordinary knack of positioning himself where he could best observe, analyze, and influence events. His influence was felt over nearly half a century, not only within the international organizations for which he worked, but among those responsible for economic and financial policy throughout the world.

What was the secret of his influence as an economic adviser? Part of the answer lies in his personality. Always a little larger than life, he was the most clubbable of men, full of good stories, parables, and telling facts, persuasive and witty, confident in simple remedies and in his power to expound them, the world's outstanding salesman of sound finance. A man so entertaining and heart warming was sure of a hearing. Added to that was an extremely well-stored mind: he took endless pains to master the factual background to financial problems, to know what was going on and put together what he could learn from others about past experience and current developments. It was not for nothing that he built up an excellent daily summary of financial news at the BIS and immersed himself so deeply in the preparation of its annual reports. When he had to suggest remedies he could start from a thorough diagnosis and take full account of special circumstances without limiting himself to purely general propositions.

His approach was neither that of the economic forecaster nor that of the pure theorist. His concern was to make sure that the right policy instruments, especially monetary instruments, were brought into use and that they operated in the right direction. Thus he was no enthusiast for fine tuning and was more given to laying emphasis on keeping costs under control than on the need for demand management. He never put together any extensive analysis

of the impact of monetary policy on output and prices but took for granted that it was at once the most powerful and most delicate of instruments available for maintaining internal and external balance.

His use of economic theory was almost always at a comparatively elementary level. But he could put his finger on the key issues and had a deep intuitive understanding of the complex interactions that govern economic development. He had courage, common sense, and conviction, so that he did not hesitate to repeat unpopular doctrine until it won acceptance. Above all, he had a very good track record. He had foreseen, well in advance, the fall in long-term interest rates in the thirties. He was never much impressed by the prediction of a prolonged dollar shortage and was one of the first to stress the growth in dollar liabilities. He also realized very early—certainly by 1942—that the postwar problem was going to be one of coping with inflation rather than with unemployment. The emphasis of what he had to say moved increasingly, therefore, with the temper of the times.

The danger that he foresaw was that cheap money and deficit spending would continue in circumstances to which they were quite inappropriate. They would issue inevitably in excess demand, external deficits, and appeals for aid from abroad when policies designed to secure "internal balance" would make such aid superfluous. Such policies should be designed to restore monetary discipline, and it was on the monetary component that he kept harping until discount rates were increased in the early fifties. The message that he sought to emphasize was the primacy of monetary policy in maintaining price stability and the need simultaneously to limit public spending and preserve budgetary balance, if not year by year at least over a period of years. This was the point at which he took issue with Keynesian thinking, and it was also at the root of his popularity with central bankers

It happens also to be the point at which he and I parted company. He had no use for the Radcliffe Committee Report of which I was a signatory. He found convincing evidence of the importance of credit control in Germany's experience in 1950–51, when an increase in the discount rate from 4 to 6 percent was followed by a gradual but pronounced swing of the German balance of payments into surplus. But this still seems to me a special case in which, for obvious reasons, stock building of imported materials had been carried to excess, and it was only a question of time before the steep upward trend in German exports predominated over speculative excesses. In general, the early postwar period was dominated by the logistical difficulties involved in regaining prewar levels of production and trade, and cheap money was not a universally important handicap in the restoration of internal balance.

However all that may be, there was no doubt about Jacobsson's priorities. He put price stability first; and since he shared neither Keynes's doubts about the self-regulating character of the economy nor Keynes's optimism about managing the level of activity with success, he was content to let employment look after itself. In some respects, therefore, he came close to the monetarist position. But he disagreed strongly with Milton Friedman on monetary management, insisting on the need for flexibility. He was also insistent on the need for what he called "a balance between prices and costs." It is not altogether clear what he meant by this, since one might suppose that the relationship between the two is governed by the state of demand. It would seem that what he was after was the danger that profit margins might be compressed by excessive wage claims or misplaced tax policies and the need to take account of cost and supply factors before deciding on the timing or scale of any boost

to demand. He envisaged the use by governments of measures designed to influence costs directly, including what has come to be called "incomes policy," so that here again he was by no means in agreement with present-day monetarists.

Although he was doubtful whether the influence of Keynes had been on balance healthy, his economic philosophy had much in common with Keynes's. He was, like Keynes, a believer in management: both the management of money and the management of the economic system through financial mechanisms. He was no believer in controls; but, like Keynes, he saw the need to find a way of stabilizing the wage level with the agreement of all parties concerned. Although he disliked deficit spending and public works, he was not prepared to condemn them in all circumstances. Even more than Keynes, he was essentially a nineteenth-century liberal, against nationalization and government subsidies, in favor of a market economy and free trade, championing the small against the large, private enterprise against government control.

The ideas for which Jacobsson contended gained increasingly in influence over the postwar world. Yet he never succeeded in committing them systematically to paper in the form of a treatise. This he regretted deeply. His compulsive need to talk—"the daily spreading of the substance of my gifts"—took precedence over the need to construct a theoretical scaffolding for his ideas. But there is no doubt that he made the right choice and that his forte lay in economic policymaking rather than in theoretical virtuosity. He left behind in his diary and his writings an ample record of the events in which he took part and of the ideas governing his approach to them.

ALEC CAIRNCROSS

Oxford

Keynes' Monetary Thought: A Study of Its Development. By DON PATINKIN. Durham, N.C.: Duke University Press, 1976. Pp. 155. $9.75.

It is a truly amazing story. He was a renowned teacher, scholar, and writer, an influential commentator on world affairs, an expert always consulted by his government, an intellectual active in partisan politics, a successful operator in business and finance, and much more. At the age of 47, he had just published his *magnum opus*, a two-volume treatise setting forth his mature conclusions on the entire range of his scientific expertise. The eagerly awaited work excited great interest in his profession, but considerable criticism too. The author himself was dissatisfied. Even before the ink was dry, he was changing his mind and working on the sequel. This task became a corporate enterprise, engaging not only the critical wisdom of his peers but the loyal enthusiasm of one of the most remarkable concentrations of talented young scholars in academic history. Seminars and lectures quickly turned from glosses on the completed work to draft chapters of the new. Over 5 years of constant conversation and correspondence, the master, his young disciples, and selected critics, friendly and not so friendly, argued out the new doctrines. He and his disciples never doubted they were at last figuring out how the world really works, solving puzzles that had stumped their predecessors for a century. They did not rely on experimental findings, statistical inference, historical study, or mathematical innovation, but on sheer logic and insight. Their

parochial confidence that truths worth knowing arose inside a circle of a 40-mile radius would have been ridiculous arrogance had it not been so nearly justified. Finally, the new book appeared and wrought the revolution its makers anticipated, not only commanding the attention of the profession worldwide for decades to come, but profoundly influencing the policies and politics of nations.

The two books were, of course, Keynes's *Treatise on Money* (1930) and his *General Theory* (1936). The publication of *The Collected Writings of John Maynard Keynes* by the Royal Economic Society makes available many details of the story, and more are accessible in unpublished documents among the Keynes Papers in the Marshall Library at Cambridge. It is the good fortune of the profession that a great monetary theorist, Don Patinkin, has undertaken active research in the history of monetary theory and in particular on Keynes. In the brief volume under review he traces Keynes's monetary and macroeconomic thought through his three major works, beginning with the *Tract on Monetary Reform* (1923), stressing the story sketched above of the two later books. Keynes's multiple roles in science and public life make for an engrossing case study of the interactions of theoretical developments with the events, problems, and policy controversies of the times. The intellectual and personal relationships of Keynes to his older colleagues (Pigou, Robertson, Hawtrey, Hayek), to his young disciples at Cambridge (Kahn, Joan Robinson, Austin Robinson, Sraffa) and at Oxford (Harrod, Meade), and to others (including Kaldor, Lerner, Myrdal, Ohlin, Hicks) are a fascinating episode in the sociology of scholarship. Patinkin's concise narratives whet the appetite, and he refers us to works by Winch, Moggridge, and Howson for full meals.

Here Patinkin's main emphasis is on theory. He clearly explains and compares propositions of the three books and related articles, and he seeks reasons for the changes. Most readers will be particularly interested in Patinkin's views of what Keynes "really meant" in the *General Theory*. With admirable restraint, Patinkin refrains from interpreting or appraising Keynes from the hindsight of today's theory and knowledge, even his own.

The *Tract* Patinkin finds fascinating on the national and international problems and policies of the era, but in monetary economics he finds it a routine recital of the Cambridge version of the quantity theory. The *Treatise* he finds dull and mechanical—"a Keynes out of character, a Keynes attempting to act the role of a Professor, and a Germanic one at that" (p. 24). The famous Fundamental Equations are fundamentally flawed, and the objective of explaining output fluctuations is bound to elude a theory that is designed to explain prices and profits for given output but lacks equations for output determination. It is this gap that led Keynes to the theory of effective demand, in Patinkin's view the central insight of the *General Theory*. For the *General Theory* and for Keynes's lifetime contribution Patinkin is a great enthusiast. In these days, when Keynes's standing with the profession and with the general public is cyclically low, when the man's writings and influence are blamed for all the world's ills, students and young economists will gain valuable perspective from Patinkin's book.

On the main points of interpretation I agree with Patinkin. (1) The pretension to establish an equilibrium with involuntary unemployment, which puts off so many economists instilled with neoclassical instincts, should not be taken seriously. Keynes is really describing an economy in disequilibrium, or in a succession of constrained temporary equilibria, using comparative statics as the analytical language of the day. His strong conviction and his true

message are that the automatic self-righting mechanisms of whole economies are slow, weak, and unreliable. Excess supply can persist for long periods, whether one calls the situation equilibrium or not. (2) Keynes did not assume either money wage rigidity or money illusion. He did contend that wages would adjust downward slowly—more slowly than prices could move upward. And he was skeptical of the employment effects of economy-wide wage reduction. As Patinkin says, he regarded wage and price deflation as the equivalent of monetary expansion. Neither might be capable of lowering interest rates enough to revive investment in depressed times. (3) Keynes emphasizes nonpolicy disturbances to the economy, real as well as monetary. The "state of long-term expectation" by entrepreneurs and investors is a major determinant of effective demand. Since these expectations relate to essentially unknowable and nonprobabilistic future events, including the expectations of future entrepreneurs and investors, they are arbitrary and volatile. Keynes also stresses, as in the *Treatise*, the importance of differences of view between savers and investors, or bulls and bears, or lenders and borrowers. (4) The Hicksian IS/LM apparatus captures beautifully the general equilibrium character of Keynes's vision, but its concise formalism does not do justice to the above insights and others. (In early drafts of the book and in lectures Keynes wrote symbols for expectations or "state of the news" as arguments of functions, but they were not in the book for Hicks to find.)

My differences from Patinkin's observations are minor. (1) He describes (p. 81) the financial side of the *General Theory* as a three-asset model (money, bonds, equities), in contrast to the two-asset model (money, equities) of the *Treatise*. This seems to be stretching things. It seems to me that the *General Theory* really has only two assets (money, everything else). Anyway, there is only one endogenous interest rate, that to which the estimated marginal efficiency of capital is to be equated by variation of investment—a stock-flow confusion, as Lerner observed early on. In treating the portfolio choice between equities and fixed-money-value assets and in considering the way the market brings together "bulls" and "bears," I think the *Treatise* is superior, dealing with more interesting and important matters than speculation about bond interest rates. (2) Patinkin notes with favor (p. 36) that the *Treatise* abandoned the quantity-theory presumption, followed in the *Tract*, that monetary measures first alter the quantity of money and only then, via this medium, affect spending, profits, prices, and output. In the *Treatise* central bank operations change interest rates and thereby affect all the variables, including the quantity of money, simultaneously. The *General Theory* appears to be a step backward in this respect. (3) Patinkin allows himself to wonder (p. 110) about Keynes's omission of wealth, in distinction from capital gains, as a determinant of consumption. Keynes in fact does recognize life-cycle effects, and he would properly not wish to count planned or expected accumulations of savings as another determinant. But why does he not recognize real gains or losses on money balances? The answer is not clear. Perhaps it is because none of Keynes's theoretical work contemplates drastic or continuing one-way price change. As in other monetary theory of the day, the *quaesitum* was the determination of the equilibrium value of money, that is, the price level, not the inflation rate or any other dynamic price path. Given this mind-set, Keynes might naturally regard short-run price variation as too transient and reversible to enter consumers' reckoning of wealth. The same mind-set must be the reason that Keynes did not see inflation as a way of lowering real

interest rates even when nominal rates were close to their floor. (4) Greater stress might be placed on Keynes's uncritical acceptance of the neoclassical competitive model. By assuming that firms are price takers in auction markets rather than price setters in monopolistic competition or oligopoly, he made it harder to sustain his vision of persistent disequilibrium, with failures of coordination, communication, and adjustment. Imperfect competition was the other revolution in economics in the 1930s; one of its sites was Keynes's Cambridge, and two of its agents, Joan Robinson and Sraffa, were in his group. Yet for some mysterious reason the two revolutions were never meshed.

Keynes planned, Patinkin reports, to write a book of "Footnotes to the *General Theory*," but was never able to do so. The fourth book, Patinkin feels, would have been one of continuation and clarification of the third. Keynes had no reason to rethink the *General Theory* as he had the *Treatise*.

Which of the several Keynesian traditions today is authentic? Any or all? Patinkin does not enter this dangerous terrain. We can hope he will tell us on some other occasion. Meanwhile we can look forward to more books that combine, as this one does, fascinating narrative, economic history, history of thought, and lucid exposition of difficult theory.

JAMES TOBIN

Yale University

Studies in Macroeconomic Theory. Volume 1: Employment and Inflation. By EDMUND S. PHELPS.
New York: Academic Press, 1979. Pp. xii + 418. $19.50.

This volume reprints 17 papers by Phelps that were originally published between 1965 and 1978, together with one previously unpublished paper from 1959 and some newly written material in the form of seven brief commentaries on the origins and aims of the papers and a synthetic introductory essay laying out a theory of price, wage, and employment fluctuations. Phelps's work is deep, insightful, and technically challenging. Many of the reprinted papers are well known and have been influential in the rapid development of macroeconomic analysis in recent years.

My candidate for the most significant paper in the volume is "Phillips Curves, Inflation Expectations, and Optimal Employment over Time" (1967). In this paper Phelps introduced the natural-rate hypothesis and the associated acceleration hypothesis into macroeconomic analysis, although he modestly reports that he learned these ideas from his mentors William Fellner and Henry Wallich. The natural-rate hypothesis probably has proven to be the most important new idea in macroeconomics since 1936.

Milton Friedman's apparently independent invention of the natural-rate hypothesis in his 1967 presidential address to the American Economic Association represents a remarkable coincidence in the history of the subject. Interestingly, however, Phelps and Friedman in their original presentations drew radically different lessons from the natural-rate hypothesis. For Friedman, the natural-rate hypothesis provided another reason for believing that activist stabilization policy offers little potential benefit to offset its inherent risks. For Phelps, the natural-rate hypothesis provided a new and welcomed

challenge to the technical virtuosity of students of optimal control techniques, a reaction that reflects an essential part of his implied view, discussed further below, of the relevance of economic analysis for economic policy.

The contents of this volume mostly are concerned with two problems: the development of positive models of economic behavior that can account for observed macroeconomic fluctuations, and the development of normative criteria for macroeconomic policy based on a dynamic theory of optimal inflation. With regard to the former problem, the newly written introductory essay attempts to consolidate a theory of macroeconomic fluctuations that, as Phelps puts it, is "developed piecemeal" in the reprinted papers.

In my view, this essay fails to confirm Phelps's claim that there now exists a satisfactory theory of employment fluctuations that explains why nominal disturbances apparently create a temporary divergence between actual and expected inflation rates and thereby affect real variables in addition to prices and wages. Recent years have seen progress, to which Phelps's work has contributed, in clarifying the objectives for such a theory, but have not seen the attainment of these objectives by either Phelps or other theorists. Both of the main existing approaches to this problem—the non-market-clearing models, exemplified in this volume by the introductory essay and the paper written jointly with John Taylor, "Stabilizing Powers of Monetary Policy under Rational Expectations" (1977), and the incomplete-information models, developed most fully by Robert Lucas and Robert Barro—seem to rely either explicitly or implicitly on contrived and unconvincing impediments to Walrasian outcomes.

In Phelps's new work, specifically, the key assertion is that costs of information transmission cause delays in market-clearing price and wage adjustments. This idea, however, receives neither formal theoretical development nor empirical confirmation. Phelps's substantial contributions to the ongoing work on a theory of macroeconomic fluctuations actually have largely involved qualification and embellishment of the essential clarification embodied in the natural-rate hypothesis—namely, that a Phillips curve consistent with neoclassical economic postulates would relate divergencies of unemployment from its equilibrium, that is, "natural," level to the difference between the inflation rate and some measure of the expected inflation rate, rather than simply to the inflation rate.

One accomplishment of the newly written material is to clear up ambiguities and inconsistencies in some of the reprinted papers with regard to market-clearing assumptions. As indicated above, Phelps now prefers the non-market-clearing approach, and the new introductory essay, in contrast to some of his earlier work, assumes that both labor and product markets fail to clear. His main defense of the non-market-clearing approach is that he regards the alternative incomplete-information approach as unrealistic, because when combined with assumptions about rational expectations it implies that only unsystematic components of monetary policy affect real variables. Phelps, however, discusses no empirical evidence that supports rejection of this hypothesis. Actually, the existing empirical analysis seems to be inconclusive. Moreover, here as elsewhere, his discussion of the existing literature is impressionistic and does not include either detailed examination of alternative models or even specific references.

Phelps also refers to "the patent unrealism of the view [implicit in the incomplete-information approach] that unemployment is wholly voluntary" (p. 27). However, as many authors, including myself, have pointed out, this

criticism is either semantic or wrong. Specifically, the recently developed view of labor market transactions as involving implicit contractual arrangements, which Phelps acknowledges to be a useful contribution, explains layoffs and other apparent evidence of nonwage rationing of employment without invoking a failure of labor markets to clear.

Phelps alludes in a few places to what seems to me to be the most telling and fundamental problem with the incomplete-information approach, incorporating rational expectations and market clearing, to modeling macroeconomic fluctuations. This problem is that the assumptions of this theory apparently imply that perceived monetary disturbances would not affect real variables, a hypothesis that seems inconsistent with the apparent relation between published monetary data and measures of output and employment. To get around this problem, the Lucas and Barro formulations of this theory implicitly employ the contrivance that rational individuals disregard current monetary data.

With regard to the implicitly contractual view of labor market transactions referred to above, Phelps emphasizes a variant, developed in a short paper written jointly with Guillermo Calvo, "Employment Contingent Wage Contracts" (1977), that assumes that employers know more about realized states of nature than do their workers. The significance of this assumption, however, seems questionable because, contrary to what Calvo and Phelps assume, employers in competitive labor markets probably would not find it worthwhile to exploit such an informational advantage. In any event, even with the Calvo-Phelps assumption about firm behavior, the contractual model does not imply that employment fluctuations should be related to published monetary data. Unfortunately, Calvo and Phelps obscure this point by implicitly introducing another questionable assumption, that the monetary authority has superior information about real and velocity shocks.

In addition to macroeconomic fluctuations, the other main focus of this volume is, as indicated above, a theory of optimal inflation. Phelps is interested in prescribing inflation policy both in the context of a deterministic steady state and in stochastic and non-steady-state contexts in which optimal control techniques become relevant. The analysis of these issues is comprehensive and introduces many subtleties and complexities. Examples of fundamental insights are the viewing of the problem of optimal inflation as involving the choice of an optimal tax mix and as being parallel to the problem of optimal growth. Importantly, Phelps clearly distinguishes and explains the respective roles of empirical knowledge and ethical postulates in prescribing policy.

It is worth recalling in this context that Phelps rejects models of macroeconomic fluctuations that assume market clearing and rational expectations. Such models imply that feedback control is irrelevant. In his more recent papers, however, Phelps takes careful account of how the introduction of rational expectations in place of adaptive expectations in non-market-clearing models radically changes the control problem. In "Disinflation without Recession: Adaptive Guideposts and Monetary Policy" (1978), he devises a subtle counterinflationary plan involving "*a dynamic monetary program* which is calculated to maintain full employment insofar as the average wage approximates its warranted path, and the institution of some form of *indicative wage planning* to guide and promote the growth of wages along that warranted path" (p. 256). In advocating this plan, he also carefully assesses the questions about economic behavior that make its success less than certain.

Phelps's analysis of optimal macroeconomic policy involves challenging intellectual exercise and undoubtedly enhances understanding of important economic relations. His discussions of policy issues, however, both in this volume and in other writings, also illustrate the confusion prevailing among us academic economists about the relevance of normative economic analysis for actual policymaking in real economies. In his critique, "The 1972 Report of the President's Council of Economic Advisers [CEA]: Economics and Government," published in the September 1972 issue of the *American Economic Review* (vol. 62, pp. 533–39) and referred to but not reprinted in the present volume, Phelps observes in a tone of anger and frustration "that economic analysis does not occupy a strategic place in the selection of economic policy" (p. 538). But he does not seem to be willing to entertain the possibility that the explanations for this observation may be systematic and immutable.

In criticizing the CEA, Phelps describes the so-called game plan of 1969–71 as involving both deception and self-deception, and he suggests that these symptoms are chronic characteristics of governmental inability to make effective use of scientific knowledge. In "Stopover Monetarism: Supply and Demand Factors in the 1972–74 Inflation" (1975), he attributes the record of poor monetary policy to inadequate quantitative knowledge about the economy. In his 1972 book, *Inflation Policy and Unemployment Theory* (New York: Norton), he writes of "the hope, however unwarranted, of [the cost-benefit approach to inflation planning] having an early policy impact" (p. ix). Moreover, with regard to alternative inflation policies, he acknowledges "the extraordinary difficulties of estimating costs and benefits" (p. xvii), as well as the fear that "if the lid is taken off the inflation question . . . the public will simply bungle its inflation decisions" (p. xv). At the end of this discussion, however, he succumbs to apparently wishful thinking when he writes, "I believe there could be a significant gain from adopting a less ritualistic, more 'calculating' approach to inflation policy" (p. xvi) and finally offers the advice to sink or swim when he writes, "Getting hold of our democratic policymaking potentialities is as much the point as the modest though valuable economic gain that can be expected to result" (p. xvi).

It is interesting that, in "Inflation Planning Reconsidered" (1978), Phelps discusses precautionary considerations that take account of the political inability to take appropriate action. But, he also displays unrealistic, ivory-tower faith in the importance of logic and ideas. For example, he apparently regards such constructs as "the social utility function," "the Rawlsian maximum criterion," and "the intersection of weighted individual preferences and technically feasible choices" to be potentially operational and to provide a meaningful basis for actual policymaking. In a newly written commentary, he admits to being amazed that the natural-rate hypothesis, a straightforward implication of neoclassical theorizing, turned out to be so controversial. In "Commodity-Supply Shock and Full-Employment Monetary Policy" (1978), he fancifully attributes a decision not to increase the money stock to accommodate supply shocks to Federal Reserve acceptance in 1974 of the latest, and at that date still unpublished, theoretical work implying that nonaccommodation is harmless.

Phelps clearly does not appreciate the correct perception of many laymen that much professional economic advice is based on academic fads. In the newly written material he states, "It may be that, as a methodological device, the postulate of rational expectations will prove to be irresistible—whatever

our doubts over its substantive reality" (p. 28). But does he really expect the public to take seriously prescriptions based on nothing more than professional prejudice? And how are laymen to distinguish solid advice from untested hypotheses? My guess is that attempts to apply the latest intellectual speculation to policy questions are subversive to efforts to enhance the role of fundamental economic analysis in policymaking.

HERSCHEL I. GROSSMAN

Brown University

The Political Economy of Germany in the Twentieth Century

Karl Hardach

"An excellent background source for political and economic analysts, business executives, or anyone else seeking to understand German economic conditions today." —*Library Journal*

$22.50

An Ownership Theory of the Trade Union

Donald L. Martin

Martin compares union behavior under proprietary (private property rights) and nonproprietary assumptions. The result is a richer set of testable theories that come closer to observed union behavior than may be found in the more conventional models of union behavior.

$16.50

At bookstores

University of California Press

Berkeley 94720

Economic Development and Cultural Change

the leading American journal in its field

edited by Bert F Hoselitz

is the main voice for

the viewpoint that the economic factors in development cannot be understood apart from the social, cultural, and political factors which affect the course of national development and in turn are affected by it.

offers multiple perspectives on

• innovation diffusion • manufacturing • agriculture • land use and reform • peasantry • rural development • banking, credit, and debt • • poverty • education and literacy • religion • family planning, fertility, and population • social structures • tradition and cultural values • external trade • foreign aid • international migration • food policies • health • theory and method

One year rates—Institutions $35.00; Individuals $22.00; Students (with faculty signature) $18.00.

Visa and Master Card accepted.

Please send order to The University of Chicago Press, 11030 S. Langley Av., Chicago, IL 60628.

8/80

Interpreting Economic Time Series

Thomas J. Sargent

University of Minnesota and Federal Reserve Bank of Minneapolis

This paper explores some of the implications for econometric practice of the principle that people's observed behavior will change when their constraints change. In dynamic contexts, a proper definition of people's constraints includes among them laws of motion that describe the evolution of the taxes they must pay and the prices of the goods that they buy and sell. Changes in agents' perceptions of these laws of motion (or constraints) will in general produce changes in the schedules that describe the choices they make as a function of the information that they possess. Until very recently, received dynamic econometric practice ignored this principle. The practice of dynamic econometrics should be changed so that it is consistent with the principle that people's rules of choice are influenced by their constraints. This is a substantial undertaking and involves major adjustments in the ways that we formulate, estimate, and simulate econometric models.

Introduction

This paper explores some of the implications for econometric practice of a single principle from economic theory. This principle is that people's observed behavior will change when their constraints change. In dynamic contexts, a proper definition of people's constraints in-

The views expressed here are solely mine and do not necessarily represent the views of the Federal Reserve Bank of Minneapolis or the Federal Reserve System. Many of my thoughts on the subject of this paper have been heavily influenced by numerous discussions with Lars Peter Hansen and Robert E. Lucas, Jr. The observations on Bayesian methods are in large part those of Hansen. Ian Bain made many helpful comments on an earlier draft. This paper is the text for the Mary Elizabeth Morgan Prize lecture, given at the University of Chicago in November 1979.

Journal of Political Economy, 1981, vol. 89, no. 2]
© 1981 by The University of Chicago 0022-3808/81/8902-0007$01.50

cludes among them laws of motion that describe the evolution of the taxes they must pay and the prices of the goods that they buy and sell. Changes in agents' perceptions of these laws of motion (or constraints) will in general produce changes in the schedules that describe the choices they make as a function of the information that they possess. Until very recently, received dynamic econometric practice ignored this principle and routinely deduced policy conclusions by assuming that people's rules of choice would not vary, for example, with the government's choices of laws of motion for variables such as tax rates, government purchases, and so on. These variables are supposed to have their effects precisely because they influence the constraints of some private agents.

The practice of dynamic econometrics should be changed so that it is consistent with the principle that people's rules of choice are influenced by their constraints. This is a substantial undertaking and involves major adjustments in the ways that we formulate, estimate, and simulate econometric models. Foremost, we need a stricter definition of the class of parameters that can be regarded as "structural." The body of doctrine associated with the "simultaneous equations" model in econometrics properly directs the attention of the researcher beyond reduced-form parameters to the parameters of "structural equations," which presumably describe those aspects of the behavior of people that prevail across a range of hypothetical environments. Estimates of the parameters of structural equations are needed in order to analyze an interesting class of policy interventions. Most often, however, included in a prominent way among the "structural equations" have been equations describing the rules of choice for private agents. Consumption functions, investment schedules, demand functions for assets, and agricultural supply functions are all examples of such rules of choice. In dynamic settings, regarding the parameters of these rules of choice as structural or invariant under interventions violates our simple principle from economic theory.

This paper describes methods for interpreting economic time series in a manner consistent with the principle that people's constraints influence their behavior. For the most part, I shall restrict things so that the dynamic economic theory is of the equilibrium variety, with optimizing agents and cleared markets. However, many of the principles described here will pertain to other types of dynamic economic theories, such as "disequilibrium" models with optimizing agents. The line of work I shall describe has diverse antecedents, of which major ones are contributions of Muth (1960, 1961), Nerlove (1967), Lucas and Prescott (1971), Telser and Graves (1971), and Lucas (1972*b*,

1976).[1] The works of Granger (1969) and Sims (1972) have provided key technical econometric foundations.

The basic idea is to interpret a collection of economic time series as resulting from the choices of private agents interacting in markets assumed to be organized along well-specified lines. The private agents are assumed to face nontrivial dynamic and stochastic optimization problems. This is an attractive assumption because the solutions of such problems are known to imply that the chosen variables (e.g., stocks of factors of production or financial assets) can exhibit serial correlation and cross-serial correlation. Since time series of economic data usually have the properties of high own-serial correlation and various patterns of cross-serial correlation, it seems that there is potential for specifying dynamic preferences, technologies, constraints, and rules of the market game that roughly reproduce the serial correlation and cross-correlation patterns in a given collection of time series measuring market outcomes. If this can be done in such a fashion that the free parameters of preferences, technologies, and constraints are identifiable econometrically, it is then possible to interpret the collection of time series as the outcome of a well-specified dynamic, stochastic equilibrium model. This paper is intended as a nontechnical summary of some of the econometric and theoretical issues involved in interpreting data in this way.

But why should anybody want to interpret time-series data as representing the results of interactions of private agents' optimizing choices? The answer is not that this way of modeling is aesthetically pleasing, although it is, nor that modeling in this way guarantees an analysis that implies no role for government intervention, which it does not. The reason for interpreting time series in this way is practical: potentially it offers the analyst the ability to predict how agents' behavior and the random behavior of market-determined variables will each change when there are policy interventions or other changes in the environment that alter some of the agents' dynamic constraints. There is a general presumption that private agents' behavior and the random behavior of market outcomes both will change whenever agents' constraints change, as when policy interventions or other changes in the environment occur. The most that can be hoped for is that the parameters of agents' preferences and technologies will not change in the face of such changes in the environment. If the dynamic

[1] Examples of work in the general line are Holt et al. (1960), Craine (1975), Crawford (1975), Geweke (1977), Sargent (1977, 1978), Blanco (1978), Hansen and Sargent (1979, 1980*a*, 1980*b*), Taylor (1979, 1980), Huntzinger (1979), Kennan (1979), Meese (1979), and Nerlove, Grether, and Carvalho (1979). The philosophy of this work is reviewed by Lucas and Sargent (1978, 1980).

econometric model is formulated explicitly in terms of the parameters of preferences, technologies, and constraints, it will in principle be possible for the analyst to predict the effects on observed behavior of changes in the stochastic environment.

Past dynamic econometric studies should usually be regarded as having been directed at providing ways of summarizing the observed behavior of interrelated variables, without attempting to infer the objectives, opportunities, and constraints of the agents whose decisions determine those variables. Most existing studies can be viewed, at best, as having estimated parameters of agents' decision rules for setting chosen variables as functions of the information they possess. Most of the better studies of consumption, investment, asset demand, and agricultural supply functions must be interpreted as having estimated such decision rules. Dynamic economic theory implies that these decision rules cannot be expected to remain invariant in the face of policy interventions that take the form of changes in some of the constraints facing agents. This means that there is a theoretical presumption that historical econometric estimates of such decision rules will provide poor predictions about behavior in a hypothetically new environment. This was Lucas's (1976) critique of econometric policy evaluation procedures as they existed in 1973.

Some readers of Lucas (1976) have interpreted the message as a call to evaluate policies by using existing econometric models differently.[2] However, one implication of Lucas's argument, and of dynamic economic theory generally, is that the formulation, identification, and estimation of the models must each be approached in substantially new and different ways. Most existing models simply cannot be saved by simulating them a little more shrewdly.[3]

[2] The papers by Anderson (1979) and Mishkin (1979) seem at least partly motivated by this interpretation

[3] The set of ideas I discuss in this paper has perhaps received most notoriety in the context of macroeconomic examples. In particular, substantial attention has been devoted to the sample economies of Lucas (1972*a*) and Sargent and Wallace (1975) in which those systematic nonneutralities that come from imputing persistently suboptimal expectations to agents were shown to disappear when the hypothesis of rational expectations was imposed on agents. Crudely put, certain classes of systematic monetary policies, in particular those which operate solely via deception, were rendered impotent in the Lucas and Sargent and Wallace examples Since the publication of these papers, many papers have been published that have described setups in which the choice of systematic policy matters, even when rational expectations prevail These papers usually invoke a source of nonneutrality not based on deception, of which there are many in standard macroeconomic theory. Papers of this class have often been interpreted as providing a defense of "pre-rational expectations" activist policies along lines that were produced by calculating optimal controls for Keynesian econometric models of the style of the late 1960s. In fact, no such defense is implied, partly because the methods by which optimal controls for government policy variables are calculated are very different in all rational expectations models from the procedures that were applied to pre-rational expectations models, but also because the ways in which

Formulating and estimating "rational expectations" models and dynamic equilibrium models of economic time series involves a variety of important conceptual and econometric issues, some of which I try to summarize in this paper. Among the issues to be treated are the following:

i) *Identification criteria.*—Prior identifying information of the Cowles Commission variety, that is, mainly exclusion restrictions, plays a much smaller role in dynamic equilibrium models. Nonlinear cross-equation restrictions implied by dynamic theory are used extensively. This shift involves important modifications of past ways of thinking about identification and estimation.

ii) *Models of error terms.*—The dynamic equilibrium modeling strategy virtually forces the researcher to think about the sources and interpretations of the error terms in the stochastic equations that he fits. The explicitly stochastic nature of the theorizing makes it difficult to "tack on" error terms after the theorizing is done, a usual procedure in the past.

iii) *The role of Granger causality.*—Granger causality turns out to be a critical concept in the formulation of dynamic economic models, as it is coincident with the condition for the appearance as an information variable in an agent's decision rule of a variable not otherwise in the agent's criterion function or constraints.

iv) *Bayesian analysis.*—Bayesian econometric techniques provide a means of mixing prior theoretical information about parameters with information from the data. Such procedures are widely used by applied time-series econometricians, although often no formal Bayesian justification is given. Dynamic economic theory provides no justification for one widely imposed class of prior restrictions which can be viewed as restrictions directly on decision rules. Instead, dynamic economic theory suggests that prior information about agents' criterion functions and constraints is what should be used in estimation. This feature of dynamic economic theory has implications for the proper implementation both of formal Bayesian procedures and of less formal procedures for constraining parameter estimates.

I shall organize my discussion around an example, namely, a

---

econometric estimates are to be constructed for rational expectations models, with or without neutralities, differ substantially from the methods applied to the Keynesian models of the 1960s The main point of the Lucas (1972*a*) and Sargent and Wallace (1975) examples is that substituting the assumption of rational expectations for "adaptive" expectations makes a critical difference for the methods both by which we should evaluate and optimally choose government policies. That same message is present in the papers of Fischer (1977), Phelps and Taylor (1977), and Hall (1978), even if superficially the differences in some qualitative features of the optimal policies under the two assumptions on expectations may have seemed less dramatic than in Sargent and Wallace's example or Lucas's.

linear-quadratic version of Lucas and Prescott's (1971) model of investment under uncertainty. I shall use this example for discussing the econometric implications of dynamic equilibrium models. I have adopted a linear-quadratic setup because it simplifies both the theoretical and econometric discussions, while illustrating many of the salient methodological implications of dynamic decision theory. Linear-quadratic optimum problems deliver difference equations that are linear in the variables and so match up nicely with much existing dynamic econometric theory. The reader familiar with Lucas and Prescott (1971) will recognize how the example can be generalized to incorporate more general specifications for the technologies, preferences, and constraints. That increased generality would make the econometric implications harder to extract than with the present setup, without altering the basic message.[4]

Investment under Uncertainty

This paper describes a linear-quadratic version of Lucas and Prescott's model of investment and uses it as a vehicle for expositing a variety of conceptual and econometric issues. The model describes the mutual determination over time of the capital stock, output, and market price of a single industry. The model can be generalized to handle multiple factors of production at the cost of what are really only technical complications. Similarly, the model could also be generalized to incorporate a set of industries, like the corn and hog industries, with interacting dynamics. Finally, I mention that it is straightforward to modify the model to incorporate much richer dynamics by generalizing the nature of the adjustment costs.

[4] Using the methods of discounted dynamic programming (e.g., Blackwell 1965), theoretical results establishing existence and uniqueness of equilibria and various qualitative features of the equilibria can often be obtained for "weak" or "general" assumptions, such as that utility is concave, constraint sets are convex and monotone in shift variables, and so on. Lucas and Prescott (1971) and Lucas (1978) give interesting illustrations of these methods. These techniques were also used by Sargent (1980*b*) to make some general observations on interpreting time-series correlations between Tobin's q variable and the aggregate rate of investment. However, for applied work, it is necessary to be able to calculate equilibria as a function of the free parameters of preferences and constraints, and it is highly desirable if the equilibria can be calculated easily. While for general functional forms it is in principle possible to calculate equilibria of recursive competitive models using a contraction mapping, in practice such methods are presently too expensive to use in empirical work. For this reason, for empirical work it is presently necessary to choose functional forms for which equilibria can be calculated either analytically or very quickly. Linear-quadratic specifications are one of the few such choices of convenient functional forms available. (Various versions of logarithmic specification are also sometimes tractable, e.g., Merton [1971].) A valuable treatment of recursive competitive equilibrium models with general specifications of functional forms is Prescott and Mehra (1980).

I define the following variables:

y_t = output of the representative firm;

n = number of firms in the industry, assumed constant over time;

Y_t = ny_t = total output of industry;

P_t = price of output;

D_{1t} = a $(p_1 \times 1)$ vector of random variables appearing in the industry demand schedule, $p_1 \geqslant 1$;

D_{2t} = a $(p - p_1) \times 1$ vector of random variables which help predict future values of the collection of variables D_{1t}, $p \geqslant p_1$;

$D_t = \begin{bmatrix} D_{1t} \\ D_{2t} \end{bmatrix}$;

w_t = rental rate on capital;

W_t = a $(q \times 1)$ vector whose first element is w_t; the remaining elements of W_t are variables that help predict future w_t's;

u_t = a random shock to demand;

ϵ_t = a random shock in the production function;

k_t = stock of capital of the representative firm; and

K_t = nk_t = total capital stock in industry.

The subscript t indexes the date to which the variable corresponds.

I further define the following polynomials in the lag operator L:

$$\delta_u(L) = 1 - \sum_{j=1}^{r_u} \delta_{uj} L^j,$$

where δ_{uj} is a scalar;

$$\delta_D(L) = I_p - \sum_{j=1}^{r_D} \delta_{Dj} L^j,$$

where δ_{Dj} is a $p \times p$ matrix and I_p is the $p \times p$ identity matrix;

$$\delta_w(L) = I_q - \sum_{j=1}^{r_w} \delta_{wj} L^j,$$

where δ_{wj} is a $(q \times q)$ matrix and I_q is the $(q \times q)$ identity matrix; and

$$\delta_\epsilon(L) = 1 - \sum_{j=1}^{r_\epsilon} \delta_{\epsilon j} L^j,$$

where $\delta_{\epsilon j}$ is a scalar.[5]

The industry consists of n identical competitive firms, each of which uses a single factor of production, capital, to produce a single output.

[5] I shall impose the condition that the zeroes of $\delta_\epsilon(z)$, $\delta_u(z)$, det $\delta_D(z)$ and det $\delta_w(z)$ each exceed unity in modulus. Actually, a weaker condition would suffice, namely, that the zeroes of these polynomials each exceed $\sqrt{\beta}$ in modulus, where β is the discount factor introduced below. These conditions on the zeroes are regularity conditions that assure that the infinite series calculated in eqq (14) and (19) converge

Output of the representative firm y_t is governed by

$$y_t = fk_t + n^{-1}\epsilon_t, \qquad f > 0, \tag{1}$$

where k_t is the representative firm's stock of capital at t, and ϵ_t is a random error in the technology. The firm knows $\{\epsilon_t, \epsilon_{t-1}, \ldots\}$, but does not know with certainty future values of the shock ϵ_t. The error ϵ_t is known to follow the r_ϵth-order Markov process

$$\delta_\epsilon(L)\epsilon_t = V_t^\epsilon, \tag{2}$$

where V_t^ϵ is a "fundamental" white-noise error term for ϵ_t.[6] The firm is assumed to know $\delta_\epsilon(L)$ and $E(V_t^\epsilon)^2$ with certainty.

The demand curve for output is given by[7]

$$P_t = A_0 - A_1 Y_t + A_2 D_{1t} + u_t, \qquad A_0, A_1 > 0, \tag{3}$$

where D_{1t} is a $(p_1 \times 1)$ vector of "demand shifters," A_2 is a $(1 \times p_1)$ vector of constants, and u_t is a random shock to the demand curve. The random term u_t obeys the r_uth-order Markov process

$$\delta_u(L)u_t = V_t^u, \tag{4}$$

where V_t^u is a fundamental white noise for u_t. The $(p_1 \times 1)$ vector of demand shifters D_{1t} consists of the first p_1 rows of the $p \times 1$ vector D_t, which follows the r_Dth-order vector autoregressive process

$$\delta_D(L)D_t = V_t^D, \tag{5}$$

where V_t^D is a $(p \times 1)$ vector white noise that is fundamental for the process D_t. The representative firm is assumed to know $\delta_u(L)$, $\delta_D(L)$, A_0, A_1, A_2, and the second moments of V_t^u and V_t^D with certainty.

At time t, total output is given by

$$Y_t = ny_t = fK_t + \epsilon_t. \tag{6}$$

The representative firm's problem is to choose a contingency plan for k_{t+j} to maximize the criterion

$$E_0 \sum_{t=0}^{\infty} \beta^t \left[P_t y_t - w_t k_t - \frac{d}{2}(k_{t+1} - k_t)^2 \right], \tag{7}$$

[6] An $(n \times 1)$ vector white noise v_t^x is said to be fundamental for an $(n \times 1)$ vector process x_t if the vector of one-step-ahead linear least-squares errors in predicting x_t from past x's can be written as a linear combination of n components of v_t^x.

[7] Since a simple static demand function is posited, all of the interesting dynamics of the model come from its supply side. Specifying a demand schedule with interesting dynamics would complicate the presentation but not alter the basic messages of our example. Telser and Graves (1971) analyze dynamic optimization problems in which much of the interesting dynamics come from a demand curve that is specified. Sargent (1979, chap. 16) analyzes a model of the labor market in which the dynamics are influenced by nontrivial dynamic optimization problems solved by both suppliers and demanders.

subject to k_0 given. In (7), E_t is the mathematical expectation operator, conditional on information known to the firm at time t. This information set will shortly be specified precisely. In (7), d is a positive constant. The term $(d/2)(k_{t+1} - k_t)^2$ is intended to represent the notion that there are costs internal to the firm of adjusting the capital stock and that these rise at an increasing rate with the absolute value of the change in capital. We assume that the rental on capital w_t is the first element of the $(q \times 1)$ vector random process W_t that obeys the r_wth-order vector autoregression

$$\delta_w(L)W_t = V_t^w, \tag{8}$$

where V_t^w is a $(q \times 1)$ vector white noise that is fundamental for W_t. The firm is supposed to know $\delta_w(L)$ and the second-moment matrix of V_t^w with certainty.

At time t, the firm chooses k_{t+1}, given the information that it has available at t. However, the maximization problem (7) is not yet well posed, since we have not completely spelled out the dynamic constraints with respect to which the maximization is supposed to occur. To complete the problem (7), we begin by substituting $(fk_t + n^{-1}\epsilon_t)$ for y_t, and $(A_0 - A_1fK_t - A_1\epsilon_t + A_2D_{1t} + u_t)$ for P_t to get

$$E_0 \sum_{t=0}^{\infty} \beta^t \Big[(A_0 - A_1fK_t - A_1\epsilon_t + A_2D_{1t} + u_t)(fk_t + n^{-1}\epsilon_t) - w_tk_t - \frac{d}{2}(k_{t+1} - k_t)^2 \Big]. \tag{9}$$

In order that the problem of maximizing (9) with respect to a contingency plan for $\{k_{t+j}\}$ be well posed, it is necessary to attribute to the firm precise views about the laws of motion of the random variables that it cannot control, but whose values influence the best choice of its own stocks of capital. For problem (9), these uncontrollable variables about which the representative firm cares are K_t, D_{1t}, u_t, ϵ_t, and w_t. The firm cares about the present and future behavior of the variables $(K_t, D_{1t}, u_t, \epsilon_t)$ because they influence the present and future behavior of the market price through the demand relationship $P_t = A_0 - A_1fK_t - A_1\epsilon_t + A_2D_{1t} + u_t$. The firm cares about the evolution of the rental process w_t because it influences its costs. We have already completely described our assumptions about the firm's views of the laws of motion of D_{1t}, u_t, ϵ_t, and w_t, namely, that the firm knows the Markov laws (4), (5), (2), and (8) that govern them, and at time t knows $D_t, D_{t-1}, \ldots, u_t, u_{t-1}, \ldots, \epsilon_t, \epsilon_{t-1}, \ldots$, and $W_t, W_{t-1}, \ldots$. To complete the specification requires that we specify the firm's views about the evolution of the aggregate capital stock K_t. We assume that the representative firm believes that the aggregate capital stock evolves according to the law

$$K_{t+1} = H_0 + H_w(L)W_t + H_D(L)D_t + H_\epsilon(L)\epsilon_t + H_u(L)u_t + H_1K_t, \tag{10}$$

where H_0 and H_1 are scalars and

$$H_w(L) = \sum_{j=0}^{r_w-1} H_{wj}L^j, \qquad \text{where } H_{wj} \text{ is } (1 \times q);$$

$$H_D(L) = \sum_{j=0}^{r_D-1} H_{Dj}L^j, \qquad \text{where } H_{Dj} \text{ is } (1 \times p);$$

$$H_\epsilon(L) = \sum_{j=0}^{r_\epsilon-1} H_{\epsilon j}L^j, \qquad \text{where } H_{\epsilon j} \text{ is a scalar; and}$$

$$H_u(L) = \sum_{j=0}^{r_u-1} H_{uj}L^j, \qquad \text{where } H_{uj} \text{ is a scalar.}$$

The representative firm is assumed to know all of the parameters of the linear law of motion (10) with certainty. The reason that we have chosen the form (10) as the firm's perceived law of motion for K will shortly become apparent.

With these specifications, the maximization of (9) is now well posed. Summarizing the setup, we have that the representative firm maximizes

$$E_0 \sum_{t=0}^{\infty} \beta^t \Big[(A_0 - A_1 fK_t - A_1\epsilon_t + A_2D_{1t} + u_t)(fk_t + n^{-1}\epsilon_t) - w_tk_t - \frac{d}{2}(k_{t+1} - k_t)^2\Big], \tag{9}$$

subject to the laws of motion[8]

$$K_{t+1} = H_0 + H_w(L)W_t + H_D(L)D_t + H_\epsilon(L)\epsilon_t + H_u(L)u_t + H_1K_t, \tag{10}$$

$$\delta_w(L)W_t = V_t^w, \tag{8}$$

$$\delta_u(L)u_t = V_t^u, \tag{4}$$

$$\delta_D(L)D_t = V_t^D, \tag{5}$$

$$\delta_\epsilon(L)\epsilon_t = V_t^\epsilon, \tag{2}$$

and subject to the information set at time t,[9]

[8] It would be straightforward to modify this setup to assume that the $\{W, u, \epsilon, D\}$ processes are each finite order mixed moving average, autoregressive processes. For the details, see Hansen and Sargent (1979).

[9] These variables completely characterize the "state" vector for the firm's problem.

$$\{K_t, k_t, W_t, W_{t-1}, \ldots, W_{t-r_w+1}, D_t, D_{t-1}, \ldots, D_{t-r_D+1},$$
$$\epsilon_t, \epsilon_{t-1}, \ldots, \epsilon_{t-r_\epsilon+1}, u_t, u_{t-1}, \ldots, u_{t-r_u+1}\}.$$

The firm maximizes (9), taking the laws of motion (8), (4), (5), (2), and (10) as given and beyond its control. The firm is assumed to behave competitively and to act as if it has no control over the aggregate capital stock K. This is a reasonable assumption if n is large. The firm is assumed to know the δ's and H's with certainty and to know the first and second moments of the V_t's.[10] We further restrict the problem so that the solution is a linear contingency plan.[11] For this to be true, it is sufficient that the least-squares predictors of future W, D, ϵ, and u's be linear in the conditioning variables. This will be true if V_t^ϵ, V_t^u, V_t^D, and V_t^w obey normal probability laws. Alternatively, the analyst can simply assume that the industry is operating under optimal linear rules. In either case, the solution of the representative firm's problem is a linear contingency plan of the form[12]

$$\begin{aligned} k_{t+1} = h_0 &+ h_w(L)W_t + h_D(L)D_t + h_\epsilon(L)\epsilon_t \\ &+ h_u(L)u_t + h_1K_t + h_2k_t, \end{aligned} \tag{11}$$

where h_0, h_1, and h_2 are scalars, and

$$h_w(L) = \sum_{j=0}^{r_w-1} h_{wj}L^j, \qquad \text{where } h_{wj} \text{ is } (1 \times q);$$

$$h_D(L) = \sum_{j=0}^{r_D-1} h_{Dj}L^j, \qquad \text{where } h_{Dj} \text{ is } (1 \times p);$$

$$h_\epsilon(L) = \sum_{j=0}^{r_\epsilon-1} h_{\epsilon j}L^j, \qquad \text{where } h_{\epsilon j} \text{ is a scalar; and}$$

$$h_u(L) = \sum_{j=0}^{r_u-1} h_{uj}L^j, \qquad \text{where } h_{uj} \text{ is a scalar.}$$

The h's of (11) are in general functions both of the parameters in the criterion function (9), the parameters of δ_w, δ_u, δ_D, and δ_ϵ appearing in

We have in mind that the firm actually has observations on values of K, k, W, D, ϵ, and u for all dates t and earlier. It turns out that the firm's decisions are optimally a function only of the information set listed in the text.

[10] It is assumed that each of V_t^w, V_t^u, V_t^D, and V_t^ϵ is orthogonal to the information set $\{W_{t-s}, u_{t-s}, D_{t-s}, \epsilon_{t-s}, s \geq 1\}$.

[11] This is because we want the stochastic difference equations describing the behavior of the system to be linear and thereby to be readily susceptible to econometric analysis.

[12] That the solution to the problem is of this form follows from linear optimal control theory (see Kushner 1971, chap. 9; Bertsekas 1976, chap. 3; Sargent 1979, chap. 14; or Kwakernaak and Sivan 1972).

$$K_{t+1} = H_0 + H_w(L)W_t + H_D(L)D_t + H_\epsilon(L)\epsilon_t + H_u(L)u_t + H_1K_t, \tag{10}$$

where H_0 and H_1 are scalars and

$$H_w(L) = \sum_{j=0}^{r_w-1} H_{wj}L^j, \qquad \text{where } H_{wj} \text{ is } (1 \times q);$$

$$H_D(L) = \sum_{j=0}^{r_D-1} H_{Dj}L^j, \qquad \text{where } H_{Dj} \text{ is } (1 \times p);$$

$$H_\epsilon(L) = \sum_{j=0}^{r_\epsilon-1} H_{\epsilon j}L^j, \qquad \text{where } H_{\epsilon j} \text{ is a scalar; and}$$

$$H_u(L) = \sum_{j=0}^{r_u-1} H_{uj}L^j, \qquad \text{where } H_{uj} \text{ is a scalar.}$$

The representative firm is assumed to know all of the parameters of the linear law of motion (10) with certainty. The reason that we have chosen the form (10) as the firm's perceived law of motion for K will shortly become apparent.

With these specifications, the maximization of (9) is now well posed. Summarizing the setup, we have that the representative firm maximizes

$$E_0 \sum_{t=0}^{\infty} \beta^t \Big\{ (A_0 - A_1 f K_t - A_1\epsilon_t + A_2 D_{1t} + u_t)(fk_t + n^{-1}\epsilon_t) - w_t k_t - \frac{d}{2}(k_{t+1} - k_t)^2 \Big\}, \tag{9}$$

subject to the laws of motion[8]

$$K_{t+1} = H_0 + H_w(L)W_t + H_D(L)D_t + H_\epsilon(L)\epsilon_t + H_u(L)u_t + H_1K_t, \tag{10}$$

$$\delta_w(L)W_t = V_t^w, \tag{8}$$

$$\delta_u(L)u_t = V_t^u, \tag{4}$$

$$\delta_D(L)D_t = V_t^D, \tag{5}$$

$$\delta_\epsilon(L)\epsilon_t = V_t^\epsilon, \tag{2}$$

and subject to the information set at time t,[9]

[8] It would be straightforward to modify this setup to assume that the $\{W, u, \epsilon, D\}$ processes are each finite order mixed moving average, autoregressive processes. For the details, see Hansen and Sargent (1979).

[9] These variables completely characterize the "state" vector for the firm's problem.

$$\{K_t, k_t, W_t, W_{t-1}, \ldots, W_{t-r_W+1}, D_t, D_{t-1}, \ldots, D_{t-r_D+1},$$
$$\epsilon_t, \epsilon_{t-1}, \ldots, \epsilon_{t-r_\epsilon+1}, u_t, u_{t-1}, \ldots, u_{t-r_u+1}\}.$$

The firm maximizes (9), taking the laws of motion (8), (4), (5), (2), and (10) as given and beyond its control. The firm is assumed to behave competitively and to act as if it has no control over the aggregate capital stock K. This is a reasonable assumption if n is large. The firm is assumed to know the δ's and H's with certainty and to know the first and second moments of the V_t's.[10] We further restrict the problem so that the solution is a linear contingency plan.[11] For this to be true, it is sufficient that the least-squares predictors of future W, D, ϵ, and u's be linear in the conditioning variables. This will be true if V_t^ϵ, V_t^u, V_t^D, and V_t^w obey normal probability laws. Alternatively, the analyst can simply assume that the industry is operating under optimal linear rules. In either case, the solution of the representative firm's problem is a linear contingency plan of the form[12]

$$\begin{aligned} k_{t+1} = h_0 &+ h_w(L)W_t + h_D(L)D_t + h_\epsilon(L)\epsilon_t \\ &+ h_u(L)u_t + h_1K_t + h_2k_t, \end{aligned} \tag{11}$$

where h_0, h_1, and h_2 are scalars, and

$$h_w(L) = \sum_{j=0}^{r_W-1} h_{wj}L^j, \qquad \text{where } h_{wj} \text{ is } (1 \times q);$$

$$h_D(L) = \sum_{j=0}^{r_D-1} h_{Dj}L^j, \qquad \text{where } h_{Dj} \text{ is } (1 \times p);$$

$$h_\epsilon(L) = \sum_{j=0}^{r_\epsilon-1} h_{\epsilon j}L^j, \qquad \text{where } h_{\epsilon j} \text{ is a scalar; and}$$

$$h_u(L) = \sum_{j=0}^{r_u-1} h_{uj}L^j, \qquad \text{where } h_{uj} \text{ is a scalar.}$$

The h's of (11) are in general functions both of the parameters in the criterion function (9), the parameters of δ_w, δ_u, δ_D, and δ_ϵ appearing in

We have in mind that the firm actually has observations on values of K, k, W, D, ϵ, and u for all dates t and earlier. It turns out that the firm's decisions are optimally a function only of the information set listed in the text.

[10] It is assumed that each of V_t^w, V_t^u, V_t^D, and V_t^ϵ is orthogonal to the information set $\{W_{t-s}, u_{t-s}, D_{t-s}, \epsilon_{t-s}, s \geq 1\}$.

[11] This is because we want the stochastic difference equations describing the behavior of the system to be linear and thereby to be readily susceptible to econometric analysis.

[12] That the solution to the problem is of this form follows from linear optimal control theory (see Kushner 1971, chap. 9; Bertsekas 1976, chap. 3; Sargent 1979, chap. 14; or Kwakernaak and Sivan 1972).

(8), (4), (5), and (2), and the H's of the perceived law of motion for capital (10). The mapping giving the h's as functions of these other parameters is defined implicitly by standard formulas in linear optimal control theory, as exposited, for example, by Kwakernaak and Sivan (1972) and Bertsekas (1976). For present purposes, it is enough to note the existence of this mapping without exploring its nature in detail. The economic content of the mapping from the δ's, H's, and objective function parameters to the h parameters of the firm's decision rule is easy to understand, since it captures the notion that the firm's rule of choice depends on both its objective and its perceived constraints (10), (8), (4), (5), and (2).

Multiplying both sides of the firm's decision rule (11) by n and using $K_t = nk_t$ gives

$$\begin{aligned} K_{t+1} = {} & nh_0 + nh_w(L)W_t + nh_D(L)D_t + nh_\epsilon(L)\epsilon_t \\ & + nh_u(L)u_t + (nh_1 + h_2)K_t. \end{aligned} \tag{12}$$

Equation (12) is the actual law of motion for aggregate capital that results from the behavior of the representative firm. The representative firm's optimization problem in effect induces a mapping from the firm's perceived law of motion for aggregate capital (10) to the actual law of motion (12). For each possible particular perceived law of motion of the form (10), there is an implied law of motion for aggregate capital of the form (12). The notion of rational expectations is that the representative firm's perceptions of (10) are correct. In effect, a rational expectations equilibrium is a fixed point of the mapping that the representative firm's optimization problem induces from (10) to (12). Formally, we define a *rational expectations equilibrium* as a perceived law of motion (10) and an implied actual law of motion (12) which are identically equal. In a rational expectations equilibrium, firms' perceptions about the law of motion for aggregate capital turn out to be confirmed by the aggregate of the choices made by firms. Upon comparing (10) with (12) it is evident that necessary and sufficient conditions for a rational expectations equilibrium are

$$\begin{aligned} H_0 &= nh_0, \\ H_w(L) &= nh_w(L), \\ H_D(L) &= nh_D(L), \\ H_\epsilon(L) &= nh_\epsilon(L), \\ H_u(L) &= nh_u(L), \\ H_1 &= (nh_1 + h_2). \end{aligned}$$

Implicit in the above definition of a rational expectations equilibrium

are the following elements: (*a*) market clearing, (*b*) optimization of the firm's expected present value, and (*c*) correct perceptions on the part of firms of the laws of motion of variables affecting their present value but beyond their control.

We begin our analysis of the model by briefly describing aspects of the optimization problem solved by the firm. Among the first-order necessary conditions for the maximization of (9) is the following system of stochastic "Euler equations," which are derived by differentiating (9) with respect to k_t for $t = 1, 2, \ldots$:

$$\begin{aligned} \beta d k_{t+1} - d(1+\beta)k_t + d k_{t-1} &= \beta w_t \\ - \beta f(A_0 - A_1 f K_t - A_1 \epsilon_t + A_2 D_{1t} + u_t), \end{aligned} \tag{13}$$

or

$$k_{t+1} - \left(\frac{1}{\beta} + 1\right)k_t + \frac{1}{\beta}k_{t-1} = \frac{1}{d}w_t - \frac{f}{d}P_t.$$

In addition to the system of Euler equations, a transversality condition is among the first-order necessary conditions. The transversality condition can be derived by methods described in Sargent (1979). The transversality condition for the present problem in effect requires that the solution possess the property

$$\lim_{j\to\infty} E_t \beta^{t+j} k_{t+j} = 0.$$

Using the lag operator, the preceding Euler equation can be rewritten as[13]

$$\left[1 - \left(\frac{1}{\beta} + 1\right)L + \frac{1}{\beta}L^2\right]k_{t+1} = \frac{1}{d}w_t - \frac{f}{d}P_t.$$

Using the factorization

$$\left[1 - \left(\frac{1}{\beta} + 1\right)L + \frac{1}{\beta}L^2\right] = \left(1 - \frac{1}{\beta}L\right)(1 - L),$$

the above Euler equation can be written as

$$\left(1 - \frac{1}{\beta}L\right)(1 - L)k_{t+1} = \frac{1}{d}w_t - \frac{f}{d}P_t.$$

Noting that $[1 - (1/\beta)L] = -\beta^{-1}L(1 - \beta L^{-1})$ and operating on both sides of the above equation with $[-\beta^{-1}L(1 - \beta L^{-1})]^{-1}$ gives the solution[14]

[13] For a discussion of the use of lag operators in the present context, see Sargent (1979, chaps. 9 and 14).

[14] In effect, the transversality condition compels us to solve the unstable root forward in this manner.

$$(1 - L)k_{t+1} = \frac{-d^{-1}\beta L^{-1}}{1 - \beta L^{-1}} w_t + \frac{\beta f d^{-1} L^{-1}}{1 - \beta L^{-1}} P_t,$$

or, equivalently,

$$(1 - L)k_{t+1} = -d^{-1}\beta \sum_{i=0}^{\infty} \beta^i w_{t+i+1} + \beta f d^{-1} \sum_{i=0}^{\infty} \beta^i P_{t+i+1}. \qquad (14)$$

It can be verified that (14) satisfies both the Euler equations and the transversality condition. Equation (14) would give the appropriate rule for setting k_{t+1} if the firm had perfect foresight about the entire future paths of the rental w_t and the output price P_t. When the firm does not have perfect foresight, the correct decision rule can be derived by replacing the future values on the right side of (14) with the corresponding mathematical expectations conditional on information the firm does have. This leads to the decision rule[15]

$$(1 - L)k_{t+1} = -d^{-1}\beta \sum_{i=0}^{\infty} \beta^i E w_{t+i+1} \mid \Omega_t + \beta f d^{-1} \sum_{i=0}^{\infty} \beta^i E P_{t+i+1} \mid \Omega_t. \quad (15)$$

Here Ω_t is defined as the information set $\Omega_t = \{W_t, W_{t-1}, \ldots, u_t, u_{t-1}, \ldots, D_t, D_{t-1}, \ldots, \epsilon_t, \epsilon_{t-1}, \ldots, K_t\}$. The conditional mathematical expectations are assumed to be computed using the laws of motion (10), (8), (4), (5), and (2) for K, W, u, D, and ϵ, respectively, as well as the demand relationship $P_t = A_0 - A_1(fK_t + \epsilon_t) + A_2 D_{1t} + u_t$, which is used to deduce the law of motion for P_t. Once these conditional mathematical expectations are explicitly calculated in terms of the parameters of (10), (8), (4), (5), (2), and the demand curve (3), they can be substituted into equation (15) to deduce the optimum decision rule (11) for the representative firm. The decision rule (11) is linear in all of the information variables that appear on the right side. However, as the above method of calculating the parameters h of the decision rule (11) suggests, the parameters h are themselves complicated nonlinear functions of the underlying parameters of the model: the parameters A_0, A_1, A_2 of the demand curve, the parameters f and d of the technology, and the parameters $\delta_u(L)$, $\delta_\epsilon(L)$, $\delta_{u'}(L)$, and $\delta_D(L)$ of the laws of motion of the random processes given from outside the model.[16] The h's are also nonlinear functions of the H's of the law of

[15] As noted above, we shall want the relevant conditional expectations to be linear. So we shall regard the $E(\cdot \mid \Omega_t)$ that appears in (15) and elsewhere as wide-sense conditional expectations, that is, linear least-squares predictors. This amounts to restricting the firm to linear decision rules, as desired.

[16] The parameters β and n also belong in this list of underlying parameters of the model. I shall usually delete these two parameters from subsequent listings of the model's underlying parameters, though they should be understood. In some applications, the analyst may want to specify counterparts of β and n completely a priori, in which case they would not be included among the free parameters of the model over which the likelihood function or other measure of "fit" is to be maximized.

motion of aggregate capital (10), which are not given from outside but are to be determined from the analysis. The nature of these nonlinearities has been characterized by Hansen and Sargent (1980*b*) and will be alluded to further below.

Equation (15), which was derived by purely formal manipulations, has the virtue of indicating clearly that the firm has an incentive to forecast future realizations of the rental w and the output price P. As a result, any state variables that the firm sees at t, and that help predict either future P's or future w's, will appear in the firm's decision rule for k_{t+1}, given by equation (11). That the h's of (11) are nonlinear functions of the parameters $\{A_0, A_1, A_2, f, d, \beta, \delta_u, \delta_\epsilon, \delta_w, \delta_D, H_u, H_\epsilon, H_w, H_D, H_0$, and $H_1\}$ stems from the nonlinear way in which the conditional mathematical expectations of future w's and P's are functions of these parameters.

In practice, to compute a rational expectations equilibrium it is not necessary ever to calculate the right side of (15). Indeed, it is never necessary explicitly to calculate the h's that determine the decision rule (11) of the representative firm. Instead, the H's of the equilibrium law of motion for the industry can be calculated directly as follows.[17] First, multiply both sides of equation (13) by n, then use $K_t = nk_t$ and collect all terms in K on the left side to get

$$\begin{aligned} \beta d K_{t+1} - [d(1+\beta) + A_1 f^2 \beta n] K_t + d K_{t-1} &= \beta n w_t \\ - \beta n f A_0 + A_1 \beta f n \epsilon_t - \beta f n A_2 D_{1t} &- \beta f n u_t. \end{aligned} \tag{16}$$

It is of some interest that (16) is itself the Euler equation for the "social planning" problem of maximizing[18]

$$\begin{aligned} E_0 \sum_{t=0}^{\infty} \beta^t \Big\{ \Big[A_0(fK_t + \epsilon_t) - \frac{1}{2} A_1 (fK_t + \epsilon_t)^2 + (fK_t + \epsilon_t) A_2 D_{1t} \\ + (fK_t + \epsilon_t) u_t \Big] - w_t K_t - \frac{1}{2} n^{-1} d (K_{t+1} - K_t)^2 \Big\}. \end{aligned} \tag{17}$$

[17] The following argument in the text provides a way of discovering Lucas and Prescott's (1971) method of calculating the rational expectations equilibrium by formulating a fictitious social planning problem that reproduces the equilibrium. It is worth remarking that Kydland and Prescott (1977) describe a recursive method of calculating a linear rational expectations equilibrium that is applicable to our problem and is distinct from the Lucas-Prescott method upon which the discussion in the text is based. Kydland and Prescott's method successfully computes the equilibrium even in instances in which the Lucas-Prescott method breaks down. These instances occur, e.g., in which there is feedback from the industry-wide aggregate capital stock K to W or D, as would occur if lagged K's appeared as states in the Markov law for W or D. In such instances, Lucas and Prescott's social planning problem fails to reproduce the rational expectations equilibrium essentially because the fictitious planner takes into account the externality that the feedback from K to W or D constitutes.

[18] This was emphasized in a more general context by Lucas and Prescott (1971).

subject to the laws of motion (8), (4), (5), and (2) for w_t, u_t, D_{1t}, and ϵ_t[19] and subject to K_0 given.

The term in brackets is the area under the demand curve, since

$$\int_0^{Y_t} (A_0 - A_1 x + A_2 D_{1t} + u_t)dx = A_0 Y_t - \frac{1}{2} A_1 Y_t^2 + Y_t A_2 D_{1t} + Y_t u_t.$$

Thus (17) is the discounted area under the demand curve minus the total costs of production. Dividing each side of (16) by βd, the Euler equation can be written

$$\begin{aligned} K_{t+1} - \left(1 + \frac{1}{\beta} + \frac{A_1 f^2 n}{d}\right) K_t + \frac{1}{\beta} K_{t-1} \\ = \frac{n}{d} w_t - \frac{nfA_0}{d} + \frac{A_1 fn}{d} \epsilon_t \\ - d^{-1} fnA_2 D_{1t} - \frac{fn}{d} u_t. \end{aligned} \tag{18}$$

It can easily be proved that there exists a λ such that

$$[1 - (1 + \beta^{-1} + A_1 f^2 n d^{-1})L + \beta^{-1} L^2] = [1 - (\lambda\beta)^{-1} L](1 - \lambda L),$$

where $|\lambda| < 1/\sqrt{\beta}$.[20] Using $-(\lambda\beta)^{-1} L(1 - \lambda\beta L^{-1}) = [1 - (\lambda\beta)^{-1} L]$, we have that the Euler equation (18) can be written as

$$\begin{aligned} -(\lambda\beta)^{-1} L(1 - \lambda\beta L^{-1})(1 - \lambda L) K_{t+1} = \frac{-nfA_0}{d} + \frac{n}{d} w_t \\ + \frac{A_1 fn}{d} \epsilon_t - \frac{fn}{d} A_2 D_{1t} - \frac{fn}{d} u_t. \end{aligned}$$

A solution of the Euler equation that also satisfies the transversality condition for the social planning problem is

$$\begin{aligned} (1 - \lambda L) K_{t+1} = \frac{+\lambda\beta nfA_0}{d} (1 - \lambda\beta)^{-1} - \frac{n\lambda\beta}{d} \frac{L^{-1}}{1 - \lambda\beta L^{-1}} w_t \\ - \frac{A_1 fn\lambda\beta}{d} \frac{L^{-1}}{1 - \lambda\beta L^{-1}} \epsilon_t + \frac{fn\lambda\beta d^{-1} L^{-1}}{1 - \lambda\beta L^{-1}} A_2 D_{1t} \\ + \frac{fn\lambda\beta d^{-1} L^{-1}}{1 - \lambda\beta L^{-1}} u_t. \end{aligned} \tag{19}$$

Recall, for example, that $(1 - \lambda\beta L^{-1})^{-1} w_t = \sum_{j=0}^{\infty} (\lambda\beta)^j w_{t+j}$. Then it can be recognized that equation (19) is the perfect foresight solution of

[19] It can also be proved that the transversality condition for (17) imposes the same condition on the solution as does the transversality condition of the representative firm.

[20] This follows directly from the observation that if z_0 is a zero of $[1 - (1 + \beta^{-1} + A_1 f^2 n d^{-1})z + \beta^{-1} z^2]$, then so is βz_0^{-1}

the planning problem that the rational expectations competitive equilibrium implicitly solves. Thus, equation (19) expresses the aggregate capital stock K_{t+1} as a linear function of K_t and all future values of w_t, ϵ_t, D_{1t}, and u_t.

By using the methods of Hansen and Sargent (1980*b*, esp. appendix A), equation (19) can be converted to the "realizable" law for K that satisfies the Euler equations and transversality conditions, and which expresses K_{t+1} as a function only of information known at time t.[21] This involves replacing the terms w_{t+i}, ϵ_{t+i}, D_{1t+i}, and u_{t+i} in (19) by the corresponding mathematical expectations conditioned on Ω_t. The resulting equilibrium law of motion for K can be shown to be

$$\begin{aligned} K_{t+1} = H_0 &+ H_w(L)W_t + H_D(L)D_t + H_\epsilon(L)\epsilon_t \\ &+ H_u(L)u_t + H_1K_t, \end{aligned} \tag{20}$$

where

$$\begin{aligned} H_0 &= \frac{+\lambda\beta nfA_0}{d(1-\lambda\beta)}, \\ H_1 &= \lambda, \\ H_w(L) &= \frac{-n\lambda\beta}{d}\phi_w\left\{\frac{L^{-1}[I-\delta_w(\lambda\beta)^{-1}\delta_w(L)]}{1-\lambda\beta L^{-1}}\right\}, \\ H_\epsilon(L) &= \frac{-A_1fn\lambda\beta}{d}\left\{\frac{L^{-1}[1-\delta_\epsilon(\lambda\beta)^{-1}\delta_\epsilon(L)]}{1-\lambda\beta L^{-1}}\right\}, \\ H_D(L) &= +fn\lambda\beta d^{-1}A_2\phi_D\left\{\frac{L^{-1}[I-\delta_D(\lambda\beta)^{-1}\delta_D(L)]}{1-\lambda\beta L^{-1}}\right\}, \\ H_u(L) &= +d^{-1}fn\lambda\beta\left\{\frac{L^{-1}[1-\delta_u(\lambda\beta)^{-1}\delta_u(L)]}{1-\lambda\beta L^{-1}}\right\}. \end{aligned} \tag{21}$$

Here ϕ_w is a $1 \times q$ vector with 1 in the first position, followed by $(q-1)$ zeroes, and ϕ_D is a $p_1 \times p$ matrix with a $(p_1 \times p_1)$ identity matrix as the first p_1 columns and zeroes elsewhere. Notice that $w_t \equiv \phi_w W_t$ and $D_{1t} \equiv \phi_D D_t$. It is convenient at this point to recall the laws of motion assumed for w_t, u_t, ϵ_t, and D_{1t}, namely,

$$\delta_w(L)W_t = V_t^w, \tag{8}$$

$$\delta_u(L)u_t = V_t^u, \tag{4}$$

[21] Note that eq (19) satisfies the first-order necessary conditions for the optimization problem but gives the planner too much information (it is "anticipative" or "nonrealizable"). The correct solution to the problem taking the information set available to the planner into account is the solution of the first-order necessary conditions that expresses K_{t+1} as a function only of information that the planner possesses at time t. Such a solution is said to be "realizable" or "nonanticipative."

$$\delta_D(L)D_t = V_t^D, \tag{5}$$

$$\delta_\epsilon(L)\epsilon_t = V_t^\epsilon. \tag{2}$$

Equation (20) expresses the equilibrium law for the industry-wide capital stock K_{t+1} as a linear function of K_t and current and past values of W, D, ϵ, and u. Current and past values of W appear in (20) because they help predict future values of the rental rate w_t, while current and past values of D, ϵ, and u appear because they are used by agents to predict the future course of the market price P. The numbers of lagged values of W, D, ϵ, and u in (20) are $r_w - 1$, $r_D - 1$, $r_\epsilon - 1$, and $r_u - 1$, as expressions (21) can be used to show.[22] Thus, the numbers of lagged values of these "information variables" W, D, ϵ, and u in (20) are entirely inherited from the specifications of the actual laws of motion for W, D, ϵ, and u in (8), (5), (2), and (4).

Notice that the appearance of w_t and D_{1t} in the objective function of the representative firm (9) (or equivalently in the objective function of the fictitious social planner [17]) gives rise to the appearance in (20) of the entire blocks of variables W_t and D_t that help predict w and D_{1t}, respectively. Thus any variables that help predict w and D_{1t}, and which agents have information on, belong in the equilibrium law of motion for industry-wide capital. The property that the remaining variables in W (or D) help predict future values of w (or D_1) is said to be the property that the remaining variables in W (or D) *Granger cause* w (or D_1). The notion of Granger causality thus turns out to be coincident with the criterion for whether random variables that do not themselves appear in the agent's criterion function nevertheless end up in the equilibrium law of motion or decision rule, essentially because they appear in the agents' constraints as information variables that help predict variables that do appear in the criterion function. It is mainly for this reason that the concept of Granger causality has played an important role in work with rational expectations models.[23]

[22] By expanding the polynomial in L, it is possible to show that

$$\frac{L^{-1}[I - \delta(\lambda\beta)^{-1}\delta(L)]}{I - \lambda\beta L^{-1}} = \delta(\lambda\beta)^{-1}\left\{\sum_{j=0}^{r-1}\left[\sum_{k=j+1}^{r}(\lambda\beta)^{k-j-1}\delta_k\right]L^j\right\}, \tag{22}$$

where $\delta(L) = I - \Sigma_{j=1}^{r}\delta_j L^j$. Notice that the polynomial on the left side of (22) is one sided in nonnegative powers of L, despite the appearance of L^{-1}, and that it is a polynomial of order $(r - 1)$, as asserted in the text. The formula (22) can be derived by mimicking the procedures used in Hansen and Sargent (1980*a*). The same mathematical techniques used by Hansen and Sargent (1980*a*) to derive expressions like (21) or (22) were independently utilized by Futia (1979) to compute linear rational expectations equilibria. Also, without knowing of Hansen and Sargent's work, John Kennan independently derived formulas similar to (22) in a personal letter to me.

[23] From this point of view, it is irrelevant whether Granger causality is consistent with one's notion of what "true" causality is. Sims (1972) has described the relationship of the concept of Granger causality to that of strict econometric exogeneity. That a random process y fail to Granger cause x is a necessary condition for x to be strictly econometri-

Equations (20) and (21) reveal explicitly how the parameters of the equilibrium law of motion for industry-wide capital are themselves nonlinear functions of the underlying parameters $\{A_0, A_1, A_2, f, d, \beta, n, \delta_w(L), \delta_u(L), \delta_\epsilon(L), \delta_D(L)\}$. The nonlinearity has two sources. First, there is the fact that λ is a nonlinear function of β and $(A_1 f^2 n d^{-1})$ via the factorization defining λ, $[1 - (\lambda\beta)^{-1}L](1 - \lambda L) = [1 - (1 + \beta^{-1} + A_1 f^2 n d^{-1})L + \beta^{-1}L^2]$. Second, given λ, the formulas for $H_w(L), H_u(L)$, $H_\epsilon(L)$, and $H_D(L)$ in (21) are nonlinear in the parameters of $\delta_w(L)$, $\delta_u(L)$, $\delta_\epsilon(L)$, and $\delta_D(L)$. Nonlinear cross-equation restrictions of the kind illustrated by (20) and (21) are the hallmark of rational expectations models. Such cross-equation restrictions are largely absent from "pre–rational expectations" dynamic econometric models.[24] The presence of these restrictions impinges on a variety of fundamental econometric and conceptual issues, including identification, the analysis of interventions, models of "error terms," and the role of "prior information." I now turn to discussing each of these issues, using (20) and (21) as an instrument.

Analysis of Interventions

At this point, it is useful to remind ourselves of the principal reason that an economist might want to construct a dynamic econometric model of an industry along the lines of our example. It is to be able to make quantitative predictions about the effects on the industry that various hypothetical interventions or "changes in the environment" will have. In the present context, a hypothetical "intervention" or "change in the environment" means a change in one of the polynomials $\delta_w(L)$, $\delta_u(L)$, $\delta_D(L)$, or $\delta_\epsilon(L)$ that describe, respectively, the stochastic processes for W, u, D, and ϵ that impinge on the market.[25] Several interesting examples of such interventions can be given, including the following:

a) Suppose that there is a specific tax imposed on sales of the

---

cally exogenous with respect to y. For this reason, the concept of Granger causality is also useful in designing specification tests. For a discussion of the relationship between Granger causality and econometric exogeneity in the context of linear rational expectations models, see Hansen and Sargent (1980*a*).

[24] Thus Fisher wrote: "In practice, except for such covariance restrictions [across disturbances in distinct structural equations], restrictions which relate the parameters of one equation to those of one or more others are extremely rare. There is no reason in principle why such cases cannot occur, however, and it may be worthwhile devoting a very short discussion to them" (1966, p. 176).

[25] By now, this is a routine and uncontroversial definition of an intervention. Applications of the techniques of optimal control theory to the calculation of macroeconometric and microeconometric policy response functions employ precisely this concept of intervention (see, e.g., Chow 1973; Kareken, Muench, and Wallace 1973; Arzac and Wilkinson 1979; and Taylor 1979).

product. Such a specific tax can be modeled as a component of $(A_2 D_{it})$. Since the behavior of the tax through time will be described by an element of the vector Markov law $\delta_D(L)D_t = V_t^D$, changes in the rule for setting the specific tax amount to changes in one of the rows of $\delta_D(L)$.

b) Suppose that there is a specific tax on the use of the factor of production. This tax can be modeled as an addition to the rental w_t. A change in the rule for setting this tax can be modeled as a change in a row of $\delta_w(L)$.

c) Suppose there is a change in the structure of the process governing the "pretax" part of the rental. Again this can be modeled as a change in one row of $\delta_w(L)$. With a little imagination, the effects of a change in the organization of the industry[26] supplying the factor might be modeled in this way.

The model leading to (20) and (21) provides a way of predicting quantitatively the effects of such changes, once agents have caught on to them. The effect of interventions in the sense described here is to change the function (20) describing the evolution of industry capital in a way predicted by the formulas given in (21). Since interventions of this class change the law of motion (20), it is necessary to have analytic methods which use the cross-equation restrictions (21) to predict how the H's of the K-law of motion (20) will change if there is a hypothetical intervention operating on one or more of the δ's.

In order to evaluate policy interventions in this way, it is essential that the H's of (20) should not be viewed as being among the free parameters of the model. Instead, the model's free parameters are to be regarded as the deeper parameters $\{A_0, A_1, A_2, f, d, \delta_w, \delta_u, \delta_\epsilon, \delta_D\}$. The researcher needs to know these parameters in order to be able to use the formulas (21) to predict the consequences of hypothetical changes in the functions δ.[27]

From the dynamic economic theory leading to (20) and (21), it is

[26] E.g., if it becomes a cartel when before it had been competitive or noncooperative in some way.

[27] A technical qualification needs to be added at this point. In order to have a model capable of predicting effects of interventions acting on the δ's, one can sometimes get by without having uniquely identified the parameters $\{A, f, d\}$. What the researcher must identify are the parameters of the characteristic polynomial of the Euler equation (16), namely the parameters ϕ_0, ϕ_1 in $\{\beta d - [d(1 + \beta) + A_1 f^2 n\beta]L + dL^2\} = (\beta\phi_1 + \phi_0 L + \phi_1 L^2)$, where $\phi_1 = d$ and $-\phi_0 = [d(1 + \beta) + A_1 f^2 n\beta]$. The theory assumes that the ϕ_j's will be invariant with respect to interventions on the δ's. If the researcher can uniquely identify the ϕ_j's, he can proceed with econometric policy evaluation, even if he cannot uniquely identify all of (A_1, f, d). In some setups, the parameters of the characteristic polynomials of the Euler equations are identified even though only an equivalence class of the counterparts of (A_1, f, d) is identified. This is enough for econometric policy evaluation to proceed. This problem is discussed by Hansen and Sargent (1980*b*). It is technically related to the "inverse optimal control" problem (see Mosca and Zappa 1979).

evident that a given numerical version of (20), estimated from historical data, cannot be used to evaluate the consequences of arbitrary input sequences for $\{W_t\}$, $\{D_t\}$, $\{\epsilon_t\}$, and $\{u_t\}$. That is, a fixed law of motion of the form (20) with given numerical values for the H's cannot be used to investigate the consequence of arbitrarily specified numerical sequences for the W, D, ϵ, and u's. In effect, a particular version of (20) can be expected to hold up only for W, D, ϵ, and u sequences drawn from a restricted domain: namely, sequences obeying the probability laws (8), (4), (5), and (2).[28]

However, until Lucas wrote in 1976, evaluating the effects of interventions in this inappropriate way was the accepted procedure in both the macroeconometric and the microeconometric literatures. Regrettably, to this day it remains the procedure used in the overwhelming majority of analyses of policy interventions. It should be emphasized once again that from the viewpoint of the dynamic decision theory described above, the question of how agents will respond to "arbitrary sequences" of "forcing variables" W, u, ϵ, and D is not well posed. In effect, unless the researcher specifies precisely the perceived laws of motion for the "forcing variables," he has not specified the constraints subject to which decision makers are thought to be acting.[29]

Thus, in order to be able to evaluate interventions operating on the δ's, it is necessary to formulate and estimate the model in terms of the parameters of preferences (A_0, A_1, A_2), technology (f and d), and the constraints (the δ's). The argument in favor of formulating and estimating the dynamic model at the level of the deep parameters $\{A_0, A_1, A_2, d, f, \delta_w, \delta_u, \delta_\epsilon, \delta_D\}$ is in much the same spirit as the usual justification for estimating "structural" parameters rather than reduced-form parameters. As Marschak (1953) argued, the researcher wants to estimate those objects which will permit him to analyze an interesting class of changes in the environment. Dynamic economic theory has forced us to reexamine whether objects long thought to be "structural," including the parameters of decision rules such as consumption, investment, and portfolio balance schedules, are correctly taken to be invariant with respect to changes in the environment. Once agents' behavior is modeled in terms of genuinely dynamic optimization problems, it becomes apparent that the parameters of observed decision rules should not be viewed as structural (see

[28] This message is at least implicit in the work by Lucas and Prescott (1971). Gordon and Hynes (1970) made the argument in an informal way. Lucas (1976) forcefully brought the message to the attention of macroeconomists.

[29] However, some economists continue to argue that existing macroeconometric models can be used to predict the effects of such arbitrary sequences (see Friedman 1978).

Muth 1961; Lucas and Prescott 1971; Merton 1971; and Lucas 1972*a*).

The Neglect of Learning

At this point it is worthwhile to discuss a modification of the preceding kind of setup which several economists have apparently had in mind.[30] For this purpose it is sufficient to consider the problem of maximizing the social welfare criterion subject to the given laws of motion (8), (4), (5), and (2) for W_t, u_t, D_t, and ϵ_t. By relabeling and reinterpreting the variables, we can think of this as a choice problem faced by a single private agent. In posing this problem, it was assumed that the agent solving the problem knows the true values of the parameters of the objective function (17) and the true values of the polynomials in the lag operator $\delta_\epsilon(L)$, $\delta_u(L)$, $\delta_D(L)$, and $\delta_w(L)$. The observation has been made that this setup fails to incorporate a model of how the agent optimally learns about the δ's from observations on past realizations of the forcing variables ϵ, u, D, and W. Presumably, if the agent has only finite histories of observations on ϵ, u, D, and W at his disposal, then at each point in time he is uncertain about the parameters of the polynomials δ. Why not modify the preceding setup to include uncertainty about the δ's and a model of optimal learning about the δ's? There seem to be three reasons why such extensions have not as yet successfully been incorporated into rational expectations models.

The first is as follows. A general model of optimal learning about the δ's is readily available in the "Kalman filter," which can be used to model how a rational agent would use observations on $(\epsilon_t, u_t, D_t, W_t)$ to revise his prior beliefs about the δ's.[31] However, with the δ's uncertain, it is no longer possible to give closed-form formulas for the optimal decision rule in terms of what are now the posterior probability distributions over the δ's. The reason that no one has yet obtained or is likely ever to obtain such closed formulas is as follows. In deriving the closed form of the restrictions (21) for the case in which the δ's are

[30] See Friedman 1979 and Modigliani 1977.

[31] See Anderson and Moore 1979. The Kalman filter provides a model of Bayesian learning about the δ's where the initial prior and the posteriors are multivariate normal. However, as Hansen points out to me, normal posteriors for the δ's are inadmissible for dynamic models of the class described here. This is because the dynamic optimization problems we consider may be ill posed for points in the parameter space of δ's for which the zeroes of det $\delta(z)$ are less than $\sqrt{\beta}$ in modulus. Only priors and posteriors that assign zero probability to this region of the parameter space are in general admissible for our problems. This rules out multivariate normal distributions. Taking account of this admissibility constraint severely complicates the task of building a model of optimal learning about the δ's.

assumed known with certainty, the Wiener-Kolmogorov prediction formula,

$$E_t W_{t+i} = \left[\frac{\delta_w(L)^{-1}}{L^i}\right]_+ \delta_w(L) W_t,$$

was used extensively. Here $[\Sigma_{j=-\infty}^{\infty} \alpha_j L^j]_+ = \Sigma_{j=0}^{\infty} \alpha_j L^j$, so that $[\quad]_+$ means "ignore negative powers of L." The Wiener-Kolmogorov formula is equivalent with the "chain rule" of forecasting (see Shiller [1972] or Sargent [1979] for expositions). These equivalent forecasting rules are known to be correct for the case in which the δ's are known with certainty. However, as Chow (1973) has pointed out, where there is a nontrivial posterior density over the δ's, there is in general no known closed-form formula such as the above one for the i-step-ahead forecast. For example, it is not true that where δ is uncertain, the correct expression for $E_t W_{t+i}$ is given by replacing the δ_{wj}'s with their posterior means in the above formula. The fact that there is no closed-form prediction formula for sufficiently general cases implies that it is impossible to derive closed-form versions of decision rules (and hence equilibria) that correspond to (21). As we shall see, for the kind of empirical work we are advocating, it is important to have a closed form for the mapping from the parameters of the objective functions (17) and the dynamic constraints to the decision rule (20). From this viewpoint, the suggestion that one ought to build a learning mechanism into rational expectations models is not useful in suggesting practical econometric alternatives to the procedures recommended here.[32]

Another drawback with incorporating learning is that, even if one could derive the decision rules in the face of uncertain δ's, the issue would arise of how to determine the prior used to initiate the learning model for the δ's. Would it be imposed a priori or estimated? If the initial prior were to be estimated, this would substantially complicate the estimation problem and add to the number of parameters.

Finally, in many settings the Bayesian learning model implies that the posterior distributions collapse about the true δ's as time passes without limit. In such settings, even if the researcher erroneously assumes that the δ's are known with certainty when in reality agents are learning about them in an optimal way, the researcher continues to obtain consistent estimators of the underlying parameters $\{A_0, A_1,$

[32] Further, notice that if the decision rules could be calculated in closed form under uncertainty about the δ's, the resulting time-series models would have time-varying coefficients and so be nonstationary. Even if calculating the decision rules were a tractable task under uncertainty about the δ's, the loss of stationarity that it would imply might well be a price that the applied economist would not be prepared to pay even in exchange for the "greater realism" of the learning assumption.

A_2, f, d, δ_w, δ_D, δ_ϵ, $\delta_u\}$ using the methods described here and in Hansen (1979) and Hansen and Sargent (1980*a*). It does seem likely that by erroneously ignoring the phenomenon of learning about the δ's, the researcher is incorrectly calculating the asymptotic covariance matrix of his estimators. However, at present nothing is known about the nature of this error. Further, since we simply do not know how to compute optimum decision rules under the assumption that agents know the δ's with uncertainty, no consistent estimators of the underlying parameters have been proposed that incorporate agents' learning about the δ's in the optimal way, to say nothing of expressions for the associated asymptotic covariance matrices.

From the preceding considerations, I draw the conclusion that incorporating optimal Bayesian learning about the δ's on the part of agents is not a research avenue that soon promises appreciable dividends for the economist interested in applying dynamic competitive models of the sort described here.

A Model of the "Error Term"

We now derive a "dynamic supply curve" for the industry by using the industry-wide production function $Y_t = fK_t + \epsilon_t$ to eliminate K from (20) in favor of Y. Multiplying both sides of (20) by f and then adding ϵ_{t+1} to both sides gives

$$\begin{aligned} Y_{t+1} = {} & H_0 f + fH_u(L)W_t + fH_D(L)D_t \\ & + fH_\epsilon(L)\epsilon_t + fH_u(L)u_t \\ & + H_1 Y_t + \epsilon_{t+1} - H_1\epsilon_t. \end{aligned}$$

Eliminating u_t by using $u_t = P_t - A_0 + A_1 Y_t - A_2 D_{1t}$ gives

$$\begin{aligned} Y_{t+1} = {} & [H_0 f - fH_u(1)A_0] + fH_u(L)P_t + fH_u(L)W_t \\ & + [fH_D(L) - fH_u(L)A_2\phi_D]D_t \\ & + [H_1 + fH_u(L)A_1]Y_t \\ & + [1 + fH_\epsilon(L)L - H_1 L]\epsilon_{t+1}. \end{aligned} \tag{23}$$

This can be written as

$$\begin{aligned} Y_{t+1} = {} & S_0 + S_p(L)P_t + S_w(L)W_t \\ & + S_D(L)D_t + S_Y(L)Y_t + S_\epsilon(L)\epsilon_{t+1}, \end{aligned} \tag{24}$$

where

$$
\begin{aligned}
S_0 &= H_0 f - f H_u(1) A_0, \\
S_p(L) &= f H_u(L), \\
S_w(L) &= f H_w(L), \\
S_D(L) &= [f H_D(L) - f H_u(L) A_2 \phi_D], \\
S_Y(L) &= [H_1 + f H_u(L) A_1], \\
S_\epsilon(L) &= [1 + f H_\epsilon(L) L - H_1 L].
\end{aligned} \tag{25}
$$

Recall that the demand curve is

$$P_t = A_0 - A_1 Y_t + A_2 D_{1t} + u_t. \tag{3}$$

Using $\delta_\epsilon(L)\epsilon_t = V_t^\epsilon$ and $\delta_u(L)u_t = V_t^u$, we can write the supply and demand curves as

$$
\begin{aligned}
Y_{t+1} = S_0 + S_p(L)P_t + S_w(L)W_t + S_D(L)D_t \\
+ S_Y(L)Y_t + S_\epsilon(L)\delta_\epsilon(L)^{-1}V_{t+1}^\epsilon,
\end{aligned} \tag{26}
$$

$$P_t = A_0 - A_1 Y_t + A_2 D_{1t} + \delta_u(L)^{-1} V_t^u. \tag{27}$$

To discuss identification and estimation of the model, we need a theory about what is unknown to the econometrician. In constructing the model, we have taken the view that all of the variables on the right-hand side of the supply and demand curves (24) and (3) (or equivalently [26] and [27]) are known to the representative firm. Thus, from the viewpoint of private agents, (26) and (27) describe exact linear functions of the right side variables in which there are no "random errors."[33] The only tractable way that has so far been discovered of introducing random errors into (26) and (27) has been to assume that the econometrician has less information than do the private agents. The smaller information set of the econometrician leads to what from his point of view are random terms in relationships to be derived from (24) and (3) or (26) and (27). The idea is to restrict the econometrician's information set relative to that of private agents in a way both that is plausible and that leads to a tractable statistical model of the error term. I shall describe two models of the error term that can be constructed in this way.

One model results from assuming that the econometrician has time series on $\{P_t, W_t, D_t, Y_t\}$ but never observes the random processes ϵ_t and u_t. On this interpretation, ϵ_t and u_t become random terms in (24)

[33] This is a consequence of the fact that the representative firm views itself as playing a dynamic "game against nature," and so finds it optimal to use a nonrandom strategy, that is, a strategy that can be expressed as an exact function of its information variables and other state variables.

and (3) from the econometrician's viewpoint.[34] In constructing the model, we have already imposed that V^{ϵ}_{t+1} is orthogonal to all of the variables on the right side of (26) and will assume that V^{u}_{t} is orthogonal to Y_t. We can also impose that V^{u}_{t} is orthogonal to D_{1t}, if we wish,[35] although we might get by with a weaker assumption.

The second model of the error term results from assuming that the econometrician sees less of D_t and W_t than do private agents. It is convenient to postpone a detailed discussion of this second model of the error and, instead, first to discuss identification and estimation under the first model of the error term.

Identification and Estimation

With this model of the error terms, we can proceed to discuss identification and estimation. First, notice that every variable that appears on the right side of the demand schedule (27) also appears on the right side of the supply schedule (26). The dynamic economic theory leading to (26) makes the reason for this clear, since any variables that help predict future prices P will appear in the supply schedule of the representative firm. This immediately implies that any variables that help predict the demand shifters D_{1t} will appear in the supply schedule.[36] The fact that no variables on the right side of the demand curve (27) are excluded from the supply schedule (26) means that, if the supply schedule is to be identified, the source of identification must be found in restrictions of a kind different from the usual exclusion restrictions treated extensively in econometrics textbooks.[37] According to the standard "order condition" for identification, equa-

[34] This is a version of the model of the error term analyzed by Hansen and Sargent (1980*a*) and Sargent (1978).

[35] We have assumed that V^{ϵ}_{t} and V^{u}_{t} are the "innovations" or one-step-ahead errors in predicting ϵ_t and u_t on the basis of observations on all variables in agents' information set at time $t-1$ (see n. 6). This implies that V^{ϵ}_{t+1} is orthogonal to all variables on the right side of (26). If we assume that V^{u}_{t} is orthogonal to V^{ϵ}_{t}, it also implies that V^{u}_{t} is orthogonal to Y_t. Imposing that V^{u}_{t} is orthogonal to D_{1t} amounts to assuming that D_{1t} is strictly exogenous in (27), which is stronger than the Granger causality assumptions already imposed on D_{1t}, namely, that except for lagged D's, no other variables in the model Granger cause D_{1t}.

[36] There is a singular class of exceptions to this statement. In the special case that

$$ED_{1t+j} \mid \{D_{t-s}\}_{s=0}^{\infty} = ED_{1t+j} \mid \{D_{2t-s}\}_{s=0}^{\infty} \tag{28}$$

for all $j \geq 1$, D_{1t}'s will appear in the demand schedule but not in the supply schedule. The condition (28) is usually thought to be exceedingly unlikely for any economic time series $\{D_{1t}\}$.

[37] These remarks about identification should be compared with Milton Friedman's discussion (1953) of the conditions needed for "supply" and "demand" to provide a useful categorization of the factors impinging on price and output. Friedman argued that the categorization was useful to the extent that it effectively sorted forces acting on price and output into mutually exclusive categories.

tion (26) is hopelessly underidentified.[38] Thus if the parameters of the model are to be identified, sources of prior information not of the exclusion variety must be available. The main source of these restrictions in the present model is the extensive body of cross-equation restrictions embodied in equations (21) and (25). Equations (21) and (25) give the parameters of the supply schedule (26) as nonlinear functions of the parameters $\{A_0, A_1, A_2, f, d, \beta, n, \delta_w(L), \delta_D(L), \delta_u(L)$, and $\delta_\epsilon(L)\}$. In general, provided that the parameters r_D and r_w, which determine the order of the autoregressive processes for D and W, and the parameters p and q, the number of elements in the vectors D and W, respectively, are large enough, these cross-equation restrictions identify or overidentify the parameters of the model. The strength of overidentification generally increases with increases in the orders r_D and r_w and the dimensions p and q.[39]

At this point it is useful to collect together the equations comprising the model as

$$\begin{aligned} Y_{t+1} = S_0 + S_p(L)P_t + S_w(L)W_t + S_D(L)D_t \\ + S_Y(L)Y_t + S_\epsilon(L)\delta_\epsilon(L)^{-1}V^\epsilon_{t+1}, \end{aligned} \tag{26}$$

$$P_t = A_0 - A_1Y_t + A_2D_{1t} + \delta_u(L)^{-1}V^u_t, \tag{27}$$

$$\delta_w(L)W_t = V^w_t, \tag{8}$$

$$\delta_D(L)D_t = V^D_t, \tag{5}$$

where

$$\begin{aligned} S_0 &= +f^2\lambda n\beta A_0 d^{-1} - fH_u(1)A_0, \\ S_p(L) &= +f^2 n\lambda\beta d^{-1}\left\{\frac{L^{-1}[I - \delta_u(\lambda\beta)^{-1}\delta_u(L)]}{1 - \lambda\beta L^{-1}}\right\}, \\ S_w(L) &= \frac{-n\lambda\beta f}{d}\phi_w\left\{\frac{L^{-1}[I - \delta_w(\lambda\beta)^{-1}\delta_w(L)]}{1 - \lambda\beta L^{-1}}\right\}, \\ S_D(L) &= +f^2 n\lambda\beta d^{-1}A_2\phi_D\left\{\frac{L^{-1}[I - \delta_D(\lambda\beta)^{-1}\delta_D(L)]}{1 - \lambda\beta L^{-1}}\right\} \\ &\quad - f^2 d^{-1}n\lambda\beta\left\{\frac{L^{-1}[I - \delta_u(\lambda\beta)^{-1}\delta_u(L)]}{1 - \lambda\beta L^{-1}}\right\}A_2\phi_D, \end{aligned} \tag{29}$$

[38] The fact that the demand curve excludes some variables that appear in the supply schedule is due to the static specification for the demand curve. This feature of the model would not survive a variety of alterations that might plausibly be used to introduce dynamics into the demand curve. E.g., if the demand schedule involved expected future prices as arguments, all variables that help to predict future prices would appear in the econometrically operational expression for current P_t that would correspond to (27).

[39] This characteristic of identification in rational expectations models has been noted in various contexts by several authors, including Lucas (1975) and Sims (1980).

$$S_Y(L) = +f^2d^{-1}n\lambda\beta A_1\left[\frac{L^{-1}[1 - \delta_u(\lambda\beta)^{-1}\delta_u(L)]}{1 - \lambda\beta L^{-1}}\right] + \lambda,$$

$$S_\epsilon(L) = 1 - \frac{A_1 f^2 n\lambda\beta}{d}\left[\frac{1 - \delta_\epsilon(\lambda\beta)^{-1}\delta_\epsilon(L)}{1 - \lambda\beta L^{-1}}\right] - \lambda L.$$

$$[1 - (\lambda\beta)^{-1}L](1 - \lambda L) = [1 - (1 + \beta^{-1} + A_1 f^2 n d^{-1})L + \beta^{-1}L^2].$$

Equations (26), (27), (8), and (5) form a statistical model for the joint process (P_t, Y_t, W_t, D_t). The model is linear in the variables but is characterized by the extensive set of cross-equation restrictions described by (29). With the model of the error terms currently under discussion, the statistical model of the (P_t, Y_t, W_t, D_t) process has been spelled out sufficiently completely that we could write down the likelihood function for a sample (P_t, Y_t, W_t, D_t), $t = 1, \ldots, T$, assuming a normal probability density for $(V_t^w, V_t^u, V_t^p, V_t^\epsilon)$.[40] Maximum likelihood estimates of the free parameters of the model $\{A_0, A_1, A_2, f, d, \delta_D(L), \delta_w(L), \delta_u(L), \delta_\epsilon(L)\}$ could then be obtained. Computational details of such procedures are described by Sargent (1977, 1978) and Hansen and Sargent (1980*a*). From the point of view of computing the estimates, it is a great practical advantage that (29) gives a set of closed-form formulas for the cross-equation restrictions imposed by the dynamic economic theory.

Application of Bayesian Methods

The fact that for the present model of the error terms it is possible to write down a normal likelihood function means that in principle Bayesian methods are applicable. Letting θ be the list of parameters of the model and Z be the data, we have

$$f_{\text{post}}\{\theta \mid Z\} = \frac{l\{Z \mid \theta\} f_{\text{prior}}\{\theta\}}{f(Z)},$$

or

$$f_{\text{post}}\{\theta \mid Z\} = l(Z \mid \theta) f_{\text{prior}}(\theta) / \int l(Z \mid \theta) f_{\text{prior}}(\theta) d\theta, \tag{30}$$

where $f_{\text{post}}\{\theta \mid Z\}$ denotes the posterior probability density, $f(Z)$ the probability density of Z, $f_{\text{prior}}\{\theta\}$ the prior density on θ, and $l\{Z \mid \theta\}$ the likelihood function. Measures of the location and dispersion of the posterior distribution of θ can be calculated, for example, by integrating $\theta^k \cdot f_{\text{post}}\{\theta \mid Z\}$ over θ for appropriate values of k. In the Bayesian view, the role of data analysis is to trace out in as revealing a way as possible the mapping defined by (30) from the prior to the

[40] See Hansen and Sargent (1980*a*) for a discussion of the details.

posterior distribution. For such an analysis to be practical, it substantially eases matters if the mapping (30) can be characterized analytically so that, for example, posterior moments such as $\int \theta f_{\text{post}}(\theta \mid Z) d\theta$ can be calculated without the need to resort to numerical integration. Zellner (1971) and Leamer (1978) describe forms of prior densities $f_{\text{prior}}(\theta)$ that have the property that the mapping (30) is one that can be written as an analytic closed form when $l(Z \mid \theta)$ is the normal likelihood function.

In the context of dynamic economic models of the class represented by (26), (27), (8), and (5), the question of whether the mapping (30) can be characterized analytically hinges on which parameters one regards as being in the list θ about which the researcher has formulated prior information. One possibility is that θ consists of the S's of (26), the A's of (27), and δ_w, δ_u, and δ_D of (27), (8), and (5). With this interpretation of θ, then since (26), (27), (8), and (5) are linear in the S's, A's, δ_w, and δ_D, it is possible to get analytic characterizations of the mapping from $f_{\text{prior}}(\theta)$ to $f_{\text{post}}(\theta \mid Z)$. For example, Leamer (1972) and Shiller (1972) have shown how priors of various forms on the S's in (26) can tractably be mapped into posteriors, in contexts where (26) is appropriately viewed as a regression equation. In effect, Leamer (1972) and Shiller (1973) provided formal Bayesian methods for imposing restrictions on lag distributions of a general kind, examples of which had long been imposed by applied econometricians. These restrictions usually corresponded to restrictions directly on our S_j's. Predating the work of Shiller and Leamer were the restrictions on distributed lags proposed by Koyck (1954), Cagan (1956), Milton Friedman (1957), Almon (1965), and Jorgenson (1966). There was also the frequently used identifying restriction that various distributed lag weights sum to unity.[41] All of these approaches view the S's themselves as among the free parameters of the model about which the researcher can reasonably be imagined to have formed views summarized by a prior distribution.

Unfortunately, the tractability of the Leamer-Shiller approach is purchased at the cost of ignoring the essential aspects of the dynamic economic theory leading to (26). According to that theory, the S's are not free parameters but are complicated functions of the parameters $\{A_0, A_1, A_2, f, d, \beta, n, \delta_w(L), \delta_D(L), \delta_u(L), \delta_\epsilon(L)\}$. It is this list of parameters about which it seems most appropriate to expect an economist to have prior beliefs. The parameters $\{A_0, A_1, A_2, f, d\}$ are the parameters describing preferences and the technology, about which the economic theorist may have some prior beliefs. The

[41] This restriction was criticized by Lucas (1972*b*) and Sargent (1971) for essentially the same reasons given here.

economists' prior beliefs about the parameters $\{\delta_w, \delta_D, \delta_u, \delta_\epsilon\}$ are presumably on a different theoretical footing than his beliefs about $\{A_0, A_1, A_2, f, d\}$, since the former list simply characterizes the serial correlation properties of the "shift variables" about which economic theory itself suggests little, although casual general observations may suggest a presumption in favor of high serial correlation, at least in some types of variables. In any event, it is the deep parameters $\{A_0, A_1, A_2, f, d, \beta, n, \delta_w, \delta_D, \delta_u, \delta_\epsilon\}$ that must be estimated, if one is to build a model that potentially overcomes Lucas's critique of econometric policy evaluation procedures.[42]

When this list of deep parameters contains the objects of interest, Bayesian analysis using (30) becomes much less tractable. This is because the likelihood function $l(Z \mid \theta)$ becomes a very complicated function of the free parameters in θ, by virtue of the complicated nature of the cross-equation restrictions illustrated in (29). Although Bayesian analysis is still possible, the researcher will be forced to use numerical methods to characterize the mapping from the prior to the posterior given in (30). For example, for a given prior, numerical integration will have to be used to calculate the moments of the posterior distribution. My own judgment is that given current computer technology, formal Bayesian estimation procedures seem prohibitively expensive for most members of the class of dynamic models considered here. This is obviously not an objection to Bayesian methods in principle. However, I believe that the high cost attached to applying Bayesian methods correctly helps to explain why they have not yet been applied extensively to estimating rational expectations models.

A Second Model of the "Error Term"

More serious limitations on the domain of Bayesian techniques emerge if the researcher embraces a second model of the error term, which we now discuss. In the second model of the error term, it is assumed that the econometrician possesses only observations on subsets $\tilde{W}_t \subset W_t$ and $\tilde{D}_t \subset D_t$ of the information variables that private agents use to forecast future w_t's and D_{1t}'s.[43] It is assumed that these

[42] A point related to that raised in n. 29 is relevant here. Priors and posteriors that assign positive probability to points in regions for which zeroes of det $\delta(z)$ are less than $\sqrt{\beta}$ in modulus are inadmissible. This is because under such distributions, for some regions in the parameter space with positive probability, the dynamic optimum problems are not well posed. Taking this into account would substantially complicate the analysis since it would involve using mathematically less tractable distributions.

[43] This model of the error term was originally proposed by Shiller (1972) in a related but somewhat different context. The model was applied in the present context by Hansen and Sargent (1980*a*). Nerlove, Grether, and Carvalho (1979) also recommend Shiller's model of the error term.

subsets of information variables follow autoregressive processes $\tilde{\delta}_D(L)\tilde{D}_t = \tilde{V}_t^D$ and $\tilde{\delta}_w(L)\tilde{W}_t = \tilde{V}_t^w$, where $\tilde{\delta}_D(L)$ and $\tilde{\delta}_w(L)$ are polynomials in the lag operator of order $\tilde{r}_D$ and $\tilde{r}_w$, respectively. Then it turns out that the equilibrium law of motion for capital, (20), can be written in a form identical to (20), except that $W, D, \delta_D, \delta_w, H_D$, and H_w are to be replaced by the corresponding objects with tildes above them, *and* that there appears an additional random disturbance η_t on the right side of (20). The cross-equation restrictions (21) continue to characterize the objects with tildes over them.[44] The random variable η_t can be shown to be orthogonal to all of the current and lagged values of $\tilde{W}$ and $\tilde{D}$.[45] However, it turns out that η_t is in general serially correlated, with serial correlation properties that depend on the joint covariance properties of those variables in D_t and W_t that the econometrician does not have observations on. In the context of this setup, it is not even possible to write down the likelihood function without specifying details of the moments of information variables in D and W that are unobservable to the econometrician. It would seem attractive to adopt an estimation procedure that avoids the implicit theorizing about the stochastic properties of the unobserved D's and W's that an estimator using the likelihood function requires.

One such estimation strategy that exploits the orthogonality of η to $\tilde{D}$ and $\tilde{W}$, without requiring all of the added details required to write down a likelihood function, has been developed by Hansen (1979). The "generalized method of moments" estimators of Hansen have the advantage of delivering estimators of the free parameters whose desirable statistical properties do not depend on any arbitrary assumptions about the serial correlation properties of the η_t's.[46] These generalized method of moments estimators were invented precisely to handle situations in which the researcher is substantially more confident of the orthogonality conditions delivered by his theorizing than he is about the serial correlation properties of the error. These methods construct statistically consistent estimators, while avoiding the need to form the likelihood function. However, in acknowledging that he does not have enough information about the disturbances to construct the likelihood function, the researcher loses the ability to employ Bayesian methods, since knowledge of the likelihood function is essential for using Bayes's law as in (30).

[44] See Hansen and Sargent 1980*a*.

[45] Ibid.

[46] Under regularity conditions provided by Hansen (1979), the estimators of the underlying parameters are shown to be consistent and most efficient within a restricted class of estimators. Hansen's discussion of the conditions for consistency, which also has implications for the conditions for consistency of maximum likelihood estimators, is at this date the key reference on issues of statistical consistency in linear rational expectations models.

Interrelated Industries

In the Lucas-Prescott model of a single industry, state variables which help the firm predict the future prices of inputs appear in the representative firm's decision rule. The laws of motion of these input prices have been taken as given from outside the model. In actuality, the prices of these inputs are usually thought to be determined by trades in another market, one source of demand for which stems from the industry being modeled by Lucas and Prescott. If this other market is modeled explicitly, nontrivial modifications also occur in the analysis of the original industry. Thus, consider the example of a corn-hog model in which part of the output of one industry, corn, is an input into the production of the other industry, hogs. If the technology is such that hog producers have an incentive to forecast future corn prices, it follows that state variables that appear in the laws of motion for the total output and price of corn will also appear in the decision rule of the representative hog producer. Also, because the corn producer has an incentive to forecast the price of corn, which depends partly on the demand for corn from hog producers, the state variables that appear in the laws of motion for total hog output and the price of hogs will appear in the optimal decision rule of the representative corn producer. Hence, each industry inherits the state variables of the other. Furthermore, the equilibria in the two industries must be defined jointly, since the laws of motion for endogenous market-wide state variables are simultaneously determined for the two markets. There continues to be a fictitious social planning problem that a rational expectations equilibrium solves, one that is a version of an interrelated factor demand problem in which the planner is jointly optimizing the performance of both industries.[47]

These remarks indicate that the analyst will often face a hard practical decision about which dynamics he takes as given from outside the model. The internal logic of this class of models tends to propel the analyst toward a general equilibrium formulation in which the laws of motion characterizing each distinct industry are determined simultaneously with the laws of motion for all industries with which it buys and sells. In any given application, the researcher will have to choose what laws of motion he takes as given from outside the model, for the purposes of the analysis at hand.

Conclusions

Remaking dynamic econometric practice so that it is consistent with the principle that agents' constraints influence their behavior is a task

[47] These claims are proved for a particular version of a "corn-hog model" in some unpublished notes (Sargent 1980*a*).

that is far from finished. Further, properly allowing for the implications of the principle will surely require abandoning many presently received ways of interpreting data. A variety of setups can be imagined that are consistent with the principle. For example, a variety of variations of the setup of this paper can be imagined in which agents optimize but have smaller information sets than have been attributed to them here. Also, information discrepancies across classes of agents can be assumed. In many such cases, endogenous variables such as prices will play an important role in conveying information to agents. In models with dynamics as complicated as those of our examples, these variations introduce substantial analytical difficulties. To date there is very little work which investigates the econometric implications of such complications to setups like ours. With or without these complications, building a dynamic econometrics that is consistent with our simple principle from economic theory is a challenging task. It is sure to require substantial changes in the ways that applied economists interpret economic time series.

References

Almon, Shirley. "The Distributed Lag between Capital Appropriations and Expenditures." *Econometrica* 33 (January 1965): 178–96.

Anderson, Brian D. O., and Moore, John B. *Optimal Filtering*. Englewood Cliffs, N.J.: Prentice-Hall, 1979.

Anderson, Paul A. "Rational Expectations Forecasts from Nonrational Models." *J. Monetary Econ.* 5 (January 1979): 67–80.

Arzac, E. R., and Wilkinson, M. "Stabilization Policies for United States Feed Grain and Livestock Markets." *J. Econ. Dynamics and Control* 1 (February 1979): 39–58.

Bertsekas, Dimitri P. *Dynamic Programming and Stochastic Control*. New York: Academic Press, 1976.

Blackwell, David. "Discounted Dynamic Programming." *Ann. Math. Statis.* 36 (February 1965): 226–35.

Blanco, Herminio. "Investment under Uncertainty: An Empirical Analysis." Ph.D. dissertation, Univ. Chicago, 1978.

Cagan, Phillip. "The Monetary Dynamics of Hyperinflation." In *Studies in the Quantity Theory of Money*, edited by Milton Friedman. Chicago: Univ. Chicago Press, 1956.

Chow, Gregory C. "Multiperiod Predictions from Stochastic Difference Equations by Bayesian Methods." *Econometrica* 41 (January 1973): 109–18.

Craine, Roger. "Investment, Adjustment Costs, and Uncertainty." *Internat. Econ. Rev.* 16 (October 1975): 648–61.

Crawford, Robert G. "An Empirical Investigation of a Dynamic Model of Labor Turnover in U.S. Manufacturing Industries." Ph.D. dissertation, Carnegie-Mellon Univ., 1975.

Fischer, Stanley. "Long-Term Contracts, Rational Expectations, and the Optimal Money Supply Rule." *J.P.E.* 85, no. 1 (February 1977): 191–205.

Fisher, Franklin M. *The Identification Problem in Econometrics*. New York: McGraw-Hill, 1966.

Friedman, Benjamin. "Discussion." In *After the Phillips Curve: Persistence of High Inflation and High Unemployment*. Federal Reserve Bank of Boston Conference Vol. no. 19. Boston: Federal Reserve Bank, 1978.

———. "Optimal Expectations and the Extreme Information Assumptions of 'Rational Expectations' Macromodels." *J. Monetary Econ.* 5 (January 1979): 23–41.

Friedman, Milton. "The Methodology of Positive Economics." In *Essays in Positive Economics*. Chicago: Univ. Chicago Press, 1953.

———. *A Theory of the Consumption Function*. Princeton, N.J.: Princeton Univ. Press (for Nat. Bur. Econ. Res.), 1957.

Futia, Carl. "Rational Expectations in Speculative Markets." Unpublished paper, Bell Telephone Laboratories, 1979.

Geweke, John. "Wage and Price Dynamics in U.S. Manufacturing." In *New Methods in Business Cycle Research: Proceedings from a Conference*, edited by Christopher A. Sims. Minneapolis: Federal Reserve Bank, 1977.

Gordon, Donald F., and Hynes, Allan G. "On the Theory of Price Dynamics." In *Microeconomic Foundations of Employment and Inflation Theory*, by Edmund S. Phelps et al. New York: Norton, 1970.

Granger, C. W. J. "Investigating Causal Relations by Econometric Models and Cross-spectral Methods." *Econometrica* 37 (July 1969): 424–38.

Hall, Robert E. "The Macroeconomic Impact of Changes in Income Taxes in the Short and Medium Runs." *J.P.E.* 85, no. 2, pt. 2 (April 1978): S71–S85.

Hansen, Lars P. "Large Sample Properties of Generalized Method of Moment Estimators." Mimeographed. Pittsburgh: Carnegie-Mellon Univ., 1979.

Hansen, Lars P., and Sargent, Thomas J. "A Note on Wiener-Kolmogorov Prediction Formulas for Rational Expectations Models." Mimeographed. Pittsburgh: Carnegie-Mellon Univ., 1979.

———. "Formulating and Estimating Dynamic Linear Rational Expectations Models." *J. Econ. Dynamics and Control* 2, no. 1 (1980): 7–46. (*a*)

———. "Linear Rational Expectations Models for Dynamically Interrelated Variables." In *Rational Expectations and Econometric Practice*, edited by Robert E. Lucas, Jr., and Thomas J. Sargent. Minneapolis: Univ. Minnesota Press, 1980. (*b*)

Holt, Charles C.; Modigliani, Franco; Muth, John F.; and Simon, Herbert A. *Planning Production, Inventories and Work Force*. Englewood Cliffs, N.J.: Prentice-Hall, 1960.

Huntzinger, R. La Var. "Market Analysis with Rational Expectations: Theory and Estimation." *J. Econometrics* 10 (June 1979): 127–45.

Jorgenson, Dale W. "Rational Distributed Lag Functions." *Econometrica* 34 (January 1966): 135–49.

Kareken, John A.; Muench, Thomas; and Wallace, Neil. "Optimal Open Market Strategy: The Use of Information Variables." *A.E.R.* 63 (March 1973): 156–72.

Kennan, John. "The Estimation of Partial Adjustment Models with Rational Expectations." *Econometrica* 47 (November 1979): 1441–55.

Koyck, Leendert M. *Distributed Lags and Investment Analysis*. Amsterdam: North-Holland, 1954.

Kushner, Harold J. *Introduction to Stochastic Control*. New York: Holt, Rinehart, & Winston, 1971.

Kwakernaak, Huibert, and Sivan, Raphael. *Linear Optimal Control Systems*. New York: Wiley, 1972.

Kydland, Finn E., and Prescott, Edward C. "Rules Rather than Discretion: The Inconsistency of Optimal Plans." *J.P.E.* 85, no. 3 (June 1977): 473–91.

Leamer, Edward E. "A Class of Informative Priors and Distributed Lag Analysis." *Econometrica* 40 (November 1972): 1059–81.

———. *Specification Searches: Ad Hoc Inference with Nonexperimental Data.* New York: Wiley, 1978.

Lucas, Robert E., Jr. "Expectations and the Neutrality of Money." *J. Econ. Theory* 4 (April 1972): 103–24. (*a*)

———. "Econometric Testing of the Natural Rate Hypothesis." In *The Econometrics of Price Determination: Conference, October 30–31, 1970, Washington, D.C.*, edited by Otto Eckstein. Washington: Board of Governors of the Federal Reserve System, 1972. (*b*)

———. "An Equilibrium Model of the Business Cycle." *J.P.E.* 83, no. 6 (December 1975): 1113–44.

———. "Econometric Policy Evaluation: A Critique." In *The Phillips Curve and Labor Markets*, edited by Karl Brunner and Allan H. Meltzer. Carnegie-Rochester Conferences on Public Policy, vol. 1. Amsterdam: North-Holland, 1976.

———. "Asset Prices in an Exchange Economy." *Econometrica* 46 (November 1978): 1429–45.

Lucas, Robert E., Jr., and Prescott, Edward C. "Investment under Uncertainty." *Econometrica* 39 (September 1971): 659–81.

Lucas, Robert E., Jr., and Sargent, Thomas J. "After Keynesian Macroeconomics." In *After the Phillips Curve: Persistence of High Inflation and High Unemployment.* Federal Reserve Bank of Boston Conference Vol. no. 19. Boston: Federal Reserve Bank, 1978.

———. "Rational Expectations and Econometric Practice" (introductory essay). In *Rational Expectations and Econometric Practice*, edited by Robert E. Lucas, Jr., and Thomas J. Sargent. Minneapolis: Univ. Minnesota Press, 1980.

Marschak, Jacob. "Econometric Measurements for Policy and Prediction." In *Studies in Econometric Method*, edited by William C. Hood and Tjalling C. Koopmans. Cowles Foundation Monograph no. 14. New Haven, Conn.: Yale Univ. Press, 1953.

Meese, Richard. "Dynamic Factor Demand Schedules for Labor and Capital under Rational Expectations." Unpublished paper. Washington: Board of Governors of the Federal Reserve System, 1979.

Merton, Robert C. "Optimum Consumption and Portfolio Rules in a Continuous-Time Model." *J. Econ. Theory* 3 (December 1971): 373–413.

Mishkin, Frederic S. "Simulation Methodology in Macroeconomics: An Innovation Technique." *J.P.E.* 87, no. 4 (August 1979): 816–36.

Modigliani, Franco. "The Monetarist Controversy or, Should We Forsake Stabilization Policies?" *A.E.R.* 67 (March 1977): 1–19.

Mosca, Edoardo, and Zappa, Giovanni. "Consistency Conditions for the Asymptotic Innovations Representation and an Equivalent Inverse Regulation Problem." *IEEE Transactions on Automatic Control* AC-24 (June 1979): 501–3.

Muth, John F. "Optimal Properties of Exponentially Weighted Forecasts." *J. American Statis. Assoc.* 55 (June 1960): 299–306.

———. "Rational Expectations and the Theory of Price Movements." *Econometrica* 29 (July 1961): 315–35.

Nerlove, Marc. "Distributed Lags and Unobserved Components in Economic Time Series." In *Ten Economic Studies in the Tradition of Irving Fisher*, by William Fellner et al. New York: Wiley, 1967.

Nerlove, Marc; Grether, David M.; and Carvalho, José L. *Analysis of Economic Time Series. A Synthesis.* New York: Academic Press, 1979.

Phelps, Edmund S., and Taylor, John B. "Stabilizing Powers of Monetary Policy under Rational Expectations." *J.P.E.* 85, no. 1 (February 1977): 163–90.

Prescott, Edward C., and Mehra, Rajnish. "Recursive Competitive Equilibrium: The Case of Homogeneous Households." *Econometrica* 48, no. 6 (September 1980): 1365–80.

Sargent, Thomas J. "A Note on the 'Accelerationist' Controversy." *J. Money, Credit and Banking* 3 (August 1971): 721–25.

———. "The Demand for Money during Hyperinflations under Rational Expectations I." *Internat. Econ. Rev* 18 (February 1977): 59–82.

———. "Estimation of Dynamic Labor Demand Schedules under Rational Expectations." *J.P.E.* 86, no. 6 (December 1978): 1009–44.

———. *Macroeconomic Theory.* New York: Academic Press, 1979.

———. "Lecture Notes on Optimal Control and Filtering." Mimeographed. Minneapolis: Univ. Minnesota, 1980. (*a*)

———. "Tobin's q and the Rate of Investment in General Equilibrium." In *On the State of Macro-Economics*, Carnegie-Rochester Conference Series, vol. 12, edited by Karl Brunner and Allan H. Meltzer. Amsterdam: North-Holland, 1980. (*b*)

Sargent, Thomas J., and Wallace, Neil. " 'Rational' Expectations, the Optimal Monetary Instrument, and the Optimal Money Supply Rule." *J.P.E.* 83, no. 2 (April 1975). 241–54.

Shiller, Robert J. "Rational Expectations and the Structure of Interest Rates." Ph.D dissertation, Massachusetts Inst. Tech., 1972.

——— "A Distributed Lag Estimator Derived from Smoothness Priors." *Econometrica* 41 (July 1973): 775–88.

Sims, Christopher A. "Money, Income, and Causality." *A.E.R.* 62 (September 1972): 540–52.

———. "Macroeconomics and Reality." *Econometrica* 48 (January 1980): 1–48.

Taylor, John B. "Estimation and Control of a Macroeconomic Model with Rational Expectations." *Econometrica* 47 (September 1979): 1267–86.

———. "Output and Price Stability: An International Comparison." *J. Econ. Dynamics and Control* 2, no. 1 (February 1980): 109–32.

Telser, Lester G., and Graves, Robert L. *Functional Analysis in Mathematical Economics: Optimization over Infinite Horizons.* Chicago: Univ. Chicago Press, 1971.

Zellner, Arnold. *An Introduction to Bayesian Inference in Econometrics.* New York: Wiley, 1971.

Activist Monetary Policy under Rational Expectations

Peter Howitt

University of Western Ontario

The purpose of this paper is to argue that the pursuit of an activist monetary policy may make economic sense even when people's expectations are formed rationally. The paper presents a simple model in which (1) prices are costly to adjust, (2) there is uncertainty concerning the parameters affecting aggregate demand, and (3) there are positive costs of gathering and processing information. By reference to this model it can be shown that an activist monetary policy may or may not be useful in offsetting aggregate disturbances, depending upon the extent of information costs and of parameter uncertainty.

I. Introduction

By an activist monetary policy I mean a policy whereby the monetary authority permits the money supply to react systematically to information concerning aggregate disturbances. The proposition commonly known as the Sargent-Wallace (hereafter SW) proposition[1] asserts that an activist monetary policy cannot succeed in offsetting aggregate disturbances when people's expectations are formed rationally unless the monetary authority possesses information un-

An earlier version of this paper was presented at the meetings of the American Economic Association, New York, December 1977. I am indebted to Russell Boyer, Stanley Fischer, Joel Fried, Herschel Grossman, Robbie Jones, David Laidler, Robert E. Lucas, Jr., Michael Parkin, Arthur Robson, and an anonymous referee for helpful comments and criticisms, and to the Social Sciences and Humanities Research Council of Canada for financial support.

[1] See Sargent and Wallace 1975. The proposition is derived in a more general context by Barro (1976).

[*Journal of Political Economy*, 1981, vol 89, no 2]
© 1981 by The University of Chicago 0022-3808/81/8902-0005$01 50

available to private agents. It asserts that in the absence of such superior information systematic variations in the money supply will be anticipated and thus neutralized even in the short run.

The SW proposition has been criticized by previous writers. Taylor (1975) pointed out that it will take private agents time before they learn the exact nature of the monetary policy.[2] During this transition period the SW proposition is invalid. Poole (1976), Fischer (1977), and Phelps and Taylor (1977) have pointed out that even if private agents possess the same information as the monetary authority they may not be able to act upon it immediately because of contractual precommitments, in which case the SW proposition is invalid. It has also been widely acknowledged that if the arguments underlying the SW proposition were valid it would be difficult to explain the observed persistence of movements in output and employment.[3]

The present paper focuses on another objection to the SW proposition: namely, that it fails to take into account (1) the costs of gathering and processing information and (2) uncertainty concerning the structure of the economic system. People may choose not to collect and process the monetary authority's information concerning aggregate disturbances when it is costly to do so.[4] Indeed, everyday observation suggests that most people pay no attention to such information. In such circumstances people may still be forming their expectations rationally, in the sense that they are making efficient use of (costly) information, yet they are not in a position to anticipate and neutralize the effects of an activist monetary policy. In order for the SW proposition to be true the monetary authority's information must not only be available to private agents it must also be economically usable by private agents.

This qualification to the SW proposition has been noted by Barro (1976), who argued nevertheless that the monetary authority could do just as well by disseminating its information as it could by varying the money supply. However, this paper attempts to show that such a disseminating policy may be inferior to an activist policy when there is uncertainty concerning the values of parameters in the economic system because (1) people may not have enough incentive to process information, and (2) even if people do process the information there is no guarantee that this will produce the same outcome as would a properly designed monetary policy. A demonstration of this argu-

[2] A similar point was made by Benjamin Friedman (1979), who showed the similarity of rational expectations to error learning during the learning process.

[3] See, however, Lucas 1975 and Sargent 1976.

[4] This point has been stressed by Rutledge (1974), Frenkel (1975), Feige and Pearce (1976), Laidler (1978), Shiller (1978), and Friedman (1979).

ment can be made precise only in terms of a formal model, but it can be sketched in the following terms.

First, even if the information concerning aggregate disturbances is provided to people they must still incur the cost of processing it if they are to use it. But they may not find it economical to incur that cost if they think that not many other people will be using the information, even when its social value exceeds the processing cost. Consider, for example, an agent who is deciding what price to charge for the single good that he sells. Suppose he is told that the demand for money has doubled. If he thinks that all other agents are using this information he can calculate that his best course of action will be to cut his price in half, along with everyone else. The information may be quite valuable to him and to society because failure to conform in reducing his price would leave him charging a suboptimal relative price. But if he thinks that no other agent is using the information his best course of action will depend on the effects of that change in the demand for money on the demand for his product. If this effect is difficult to estimate accurately then the information may have little value to him. A convention like universal indexing that leads people to believe that everyone else will be using the information as well may be necessary for the private value of such information to be as great as its social value. Everyday observation again suggests that such a convention rarely exists.

Second, in a world without complete Arrow-Debreu contingent-claim markets there is no guarantee that even an otherwise frictionless competitive equilibrium will be efficient with respect to the information possessed by private agents. It is possible that the monetary authority is in a position to make better use of aggregate information than are private agents, even if the above-mentioned indexing convention could be established. In the example of the previous paragraph, Brainard's (1967) analysis suggests that if the values of the parameters of the economic system are not known with certainty then the optimal reaction of the monetary authority is to allow the money supply to increase, but by less than 100 percent. But if the information is disseminated prices may indeed fall in half, which would be equivalent to a 100 percent increase in the money supply. As in the theory of the second best, full information results in the best equilibrium, but when information is incomplete then more information does not necessarily result in a better equilibrium.[5]

[5] This result is similar to those of Hirshleifer (1971) and Grossman and Stiglitz (1976), who argued that the equilibrium allocations with more information may be Pareto inferior to those with less.

The remainder of the paper attempts to make these arguments more precise by developing a simple model of a monetary economy in which prices are costly to adjust. Section II describes the basic features of the model; Section III analyzes the private decision whether or not to use information; Section IV characterizes the optimal policy of the monetary authority showing how an activist policy may be justified by the first of the above arguments; Section V considers a variant of the model in which the second of the above arguments may be used to justify an activist policy; and Section VI contains some concluding remarks.

II. The Basic Model

Preliminaries

This section describes a simple hypothetical economy possessing the features described above. In this economy there are two tradable objects: money and a good. There are also two classes of economic actors: a monetary authority, whose job it is to determine how much money will exist and to distribute the money, and a set of private agents (referred to simply as "agents"), who produce and consume the good. In the interest of simplicity we suppose that each agent possesses identical tastes, technology, and initial endowments. (Thus we are making no distinction between firms and households.) In order to establish a basis for exchange among these identical agents we assume that each of them has an absolute aversion to consuming units of the good that he himself has produced. Thus he is induced to sell all of these units and to purchase for consumption units of the good that are produced by others.

To keep the demand and supply functions simple we suppose that each agent evaluates his demands according to a Stone-Geary utility function and produces subject to a quadratic cost function. That is, each agent's utility during a period depends upon $\hat{q}$, the amount of the good consumed during the period, $\hat{m}$, the real value of money balances held at the end of the period, and q, the amount of the good produced during the period, according to the particular function

$$U = K\hat{q}^{a}(\hat{m} + x)^{1-a} - (k/2)q^{2} \tag{1}$$

where k is a positive, nonstochastic parameter equal to the slope of the agent's marginal cost schedule, a and x are independent stochastic parameters with $0 < a < 1$, and K is a stochastic parameter defined as $K \equiv a^{-a}(1 - a)^{-(1-a)}$. For reasons that will become apparent later we shall refer to the parameter x as "the level of aggregate demand."

In order for purely monetary changes potentially to have real effects even in the short run we must not assume that there exists a central Walrasian auctioneer who is able each period to compute a single market-clearing price at which all trades are to occur. The actual transaction prices must be determined before the agents have collected enough information for them to be able to compute this Walrasian equilibrium price. We accomplish this by assuming that each agent is responsible for setting his own selling price. He must set the price before learning exactly how much he will be able to sell, and all trades must occur at these posted prices whether or not markets are clearing in the usual sense.[6] Assume also that the quantities transacted at these prices always equal the quantities demanded. In other words, when markets are not clearing the sellers bear all the resulting quantity rationing.[7]

At the time of the price-setting decision each agent's expected utility will depend negatively on the variance of his error in forecasting demand, because he must produce subject to increasing marginal cost. Thus he will have an incentive before setting his price to gather and process information concerning the demand function facing him so as to reduce the variance of this forecast error. The decision on what information to use at this stage will be the main focus of the following analysis.

When determining the value of M, the nominal per capita stock of money, the monetary authority is assumed to be unaware of the exact state of aggregate demand. Suppose, however, that he is able to acquire partial information by observing a variable, s, that is correlated with aggregate demand. In particular, assume that

$$x = cs \tag{2}$$

[6] This setup may be contrasted with Lucas's (1972), in which the agents are randomly divided into subgroups, each with a miniature Walrasian auctioneer, and with no communication between groups during the market period. Both setups require agents to make trading commitments before being completely informed of other agents' dispositions to trade. Lucas's approach has the advantage that market-clearing models have been more thoroughly analyzed in the literature than price-setting models. Ours has the advantage of replacing the Walrasian auctioneer with a scheme that bears some resemblance to the way prices actually get determined in many real-world markets. Whether our results could be transposed to a suitably constructed "market-clearing" model is an open question.

[7] This rule also appears to be compatible with the way transaction quantities are determined in many real-world markets, especially those organized by specialist traders—wholesalers, retailers, etc.—who post prices and hold inventories which they allow to act as buffer stocks against fluctuations in demand. By absorbing the nonprice rationing they provide a service to customers, for which they are rewarded by a spread between their buying and selling prices. In effect we are assuming that each agent is a combination consumer-producer-trader. (However, we are ignoring the potentially important role of inventories by assuming that the good is not storable.)

where c is a stochastic parameter. We shall refer to s as the "indicator" of monetary policy. The monetary authority is assumed to obey the following rule:[8]

$$M = \bar{M}(1 - gs) \tag{3}$$

where $\bar{M}$ (> 0) is the average per capita stock of money and g is the "policy reaction coefficient." Thus the monetary authority follows a "feedback" rule of reacting to observable information concerning the level of aggregate demand, and its two decision variables are $\bar{M}$ and g. The SW proposition asserts that the stochastic behavior of aggregate output in the economy will be independent of the monetary authority's choice of g.

The Sequence of Activities

During each period activities proceed according to the following sequence. First, there is a stage in which the monetary authority decides on his policy. That is, he chooses values for g and $\bar{M}$. When making this decision he is assumed to possess an accurate model of the economy, including exact knowledge of all deterministic variables, parameters, and functions, and the probability distributions of all stochastic variables. He does not, however, know the exact value that will be assumed during the period by any stochastic variable. As soon as he makes his decision all of his information is conveyed to each private agent, including the chosen values of g and $\bar{M}$. Thus there is no scope for the monetary authority to fool private agents, even in the short run, about the nature of his policy, or to make this policy decision using information that is not available to private agents.

Second, there is a stage in which private agents make their monitoring decision. That is, each one must choose whether to pay the cost of learning the exact value of s before setting his price, in which case he will be able to reduce the variance of his forecast error, or to wait until after his price has been set, when the value of s will be revealed to him at no cost. In the interest of simplicity we assume that the indicator is the only stochastic variable whose value is potentially observable before the price-setting decision is made.

Third, there is a stage in which the exact value of the indicator s is determined. Any agent who has decided to monitor observes s and pays the cost at this stage, as does the monetary authority (unless he has set $g = 0$). Note that once again we are not giving the monetary authority any informational advantage. At this stage, before he

[8] A purely random, unpredictable component could be added to (3) but, as Barro (1976) has also shown, the optimal policy, if feasible, would be to set its variance to zero.

chooses his price anyone can choose to be as well informed as the monetary authority.

The fourth stage is the one in which the values of all prices and of the money supply are decided. The money supply is determined automatically by equation (3), and the private agents choose their prices in such a way as to maximize expected utility. If they have decided to monitor the indicator (to learn the value of s), then private agents can forecast the exact value of M by using equation (3) and can use this knowledge in making their price decision. Otherwise they must make their decisions on the basis of expectations concerning M, which they form using (3) and their knowledge of the probability distribution of s. When setting his own price no agent is able to observe the others' prices. However, he is able to tell whether or not the others are monitoring the indicator, and this will be enough in most cases to allow him to predict exactly what the others' prices will be.

The fifth stage is the one in which all uncertainty is resolved. In this stage the monetary authority distributes money to the private agents in equal amounts, and the exact values of a, x, K, c, and s are revealed to everyone. Whatever values of the utility parameters a, x, and K are determined in this stage, they are shared by all agents in common. Thus, with this information each agent is able (i) to calculate exactly how much his sales will be for the period and (ii) to compute his own demands for money and the good.

In the sixth and final stage the agents visit one another to execute their planned purchases, and each agent must produce to meet the demands placed upon him. Up until this stage each of the identical agents will have been led to make identical decisions. Thus every agent will be charging the same price. Assume that in such a situation the quantity demanded from each agent will be identical.[9]

To understand what happens in this sequence we must start at the end and work back. Consider an agent who, during the sixth stage, is going to be called upon to produce the quantity q. Suppose that all other agents have set the same price, P, and that he himself has set the price rP. Then during the fifth stage he will be aware of the value of q and of the exact values of all the parameters in his utility function (eq. 1). He will formulate his own demands, $\hat{q}$ and $\hat{m}$, so as to maximize this utility function subject to the budget constraint

$$\hat{m} + \hat{q} = m + rq \tag{4}$$

[9] Alternatively, we could suppose that they all face a stochastic demand function with independent, identical probability distributions, the randomness representing the indeterminacy in the distribution of sales among identical sellers charging identical prices. However, the addition of one more random factor would add nothing but notational complexity to the analysis.

where $m = M/P$ denotes his initial holding of real balances.[10] Thus his demand for the good will be

$$\hat{q} = a(m + rq + x), \tag{5}$$

and by substituting from (4) and (5) into (1) we can see that the level of utility that he will obtain is determined by the indirect utility function

$$V(m, q, x, r) = m + rq + x - (k/2)q^2. \tag{6}$$

During the fifth stage each agent will also be able to predict his sales, q, in the following way. First, suppose that every agent is charging the same price, P. Then every agent will sell the same quantity, q, and will formulate demands according to (5) with $r = 1$. But every agent will also realize that his sales, q, will be equal to the per capita demand, $\hat{q}$. Thus the values of q and $\hat{q}$ can be computed by substituting $q = \hat{q}$ into (5), with $r = 1$, to produce

$$q = b(m + x) \tag{7}$$

with

$$b = \frac{a}{1 - a} > 0. \tag{8}$$

Equation (7) indicates the amount that each agent would demand in stage 5, as well as the amount that he would expect to sell, under the assumption that $r = 1$. And indeed the equilibrium will be one in which $r = 1$. But to see how prices are chosen we must indicate how much each agent would expect to sell if $r \neq 1$. Let us suppose that this quantity is given by the expression

$$q(m, x, b, r) = \max\,[b(m + x) + n(1 - r), 0] \tag{9}$$

where $n > 0$ is a measure of the "conjectured competitiveness" in the economy. In other words, it measures the extent to which each agent supposes that he would lose sales by raising his price above the market price. The limiting case in which $n = +\infty$ is the one usually associated with the assumption of perfect competition.

[10] It has been argued that even in markets where demanders choose quantities at predetermined prices the demand decisions must take into account the future relationships between trading partners, as well as the usual objectives, in which case the quantities chosen may be on neither the demand nor the supply curve as usually conceived (Barro 1977). This consideration is probably important in markets where exchange is conducted on a highly personal basis because of the costs of switching trading partners, as in labor markets, markets for personal credit, and so forth. However, there are also many markets where exchange is quite impersonal and buyers may, because they can remain anonymous to the sellers, choose their quantities without taking this consideration into account. When we assume that demanders choose their quantities according to the usual sort of utility maximization we are implicitly assuming this sort of impersonal arrangement.

We now have introduced a total of seven random variables: b, c, s, x, a, K, and M. Let us suppose that the probability distributions of the first three of these are mutually independent and that

$$\begin{aligned} &E(b) = \beta > 0; E(c) = \gamma > 0; E(s) = 0 \\ &\text{var}(b) = \sigma_b^2 > 0; \text{var}(c) = \sigma_c^2 > 0; \text{var}(s) = \sigma_s^2 > 0. \end{aligned} \tag{10}$$

At the outset of the fourth stage each agent realizes that if he sets the price rP his utility will equal

$$\frac{\overline{M}}{P}(1 - gs) + cs + rq\left[\frac{\overline{M}}{P}(1 - gs), cs, b, r\right] - \left(\frac{k}{2}\right)q\left[\frac{\overline{M}}{P}(1 - gs), cs, b, r\right]^2. \tag{11}$$

(This is obtained by substituting from [2], [3], and [9] into [6].) He will know that the stochastic variables b, c, and s obey (10), and he may or may not know the exact value of s, depending on whether or not he has chosen to monitor s in the second stage. Suppose that in stage 2 he has made the same monitoring decision as every other agent. Then he will be able to calculate P, the price set by all other agents. He will also know the values of $\overline{M}, g$, and k. Thus he will choose a relative price so as to maximize the expected value of (11) conditional on the information set I, where $I = \{s\}$ if he has monitored and $I = \emptyset$ otherwise. The first-order condition of this maximization (assume that $q[m, x, b, r] > 0$ *w.pr.* 1) can be expressed as

$$E\{b[\overline{M}(1 - gs)/P + cs] + n(1 - r) \mid I\} = r\bar{q}_n \tag{12}$$

where

$$\bar{q}_n = n/(1 + nk). \tag{13}$$

Let the superscript i denote the information set, $i = 1$ denoting $I = \{s\}$ and $i = 2$ denoting $I = \emptyset$. Then the solution to (12) can be written as the relative-price function

$$r = r_n^i(P, \overline{M}, g, s), i = 1, 2. \tag{14}$$

In order to tell what price to set each agent must also calculate P, the price that the other agents each will set. To do this he calculates the value of P that makes the value of r in (14) equal to unity. That is, he expects the price to be one such that every agent (including himself) is induced to conform. This will also be the actual price set by all agents in stage 4. It can be expressed as

$$P = P_n^i(\overline{M}, g, s), i = 1, 2, \tag{15}$$

and, from (12), it must satisfy the condition

$$E(q \mid I) \equiv E\{b[\overline{M}(1 - gs)/P + cs] \mid I\} = \bar{q}_n. \tag{16}$$

In the limiting case of perfect competition, (16) states that the market price will be set so that the expected market demand (per capita) equals the competitive market supply (per capita), $\bar{q} \equiv \lim_{n\to\infty} \bar{q}_n = 1/k$. In other words, there is expected market clearing.[11]

Let $\bar{m}_n \equiv \bar{q}_n/\beta$. From (10), (15), and (16) the prices can be expressed as

$$P_n^1(\bar{M}, g, s) = \bar{M}(1 - gs)/(\bar{m}_n - \gamma s), \tag{17}$$

$$P_n^2(\bar{M}, g, s) = \bar{M}/\bar{m}_n. \tag{18}$$

Thus the quantity of output that will be produced per person in stage 6 can be expressed as a function of the information set, the reaction coefficient g, and the random variables b, c, and s by substituting from (2), (3), (17), and (18) in (7) to obtain

$$q_n^1 = b[\bar{m}_n + (c - \gamma)s], \tag{19}$$

$$q_n^2 = b[\bar{m}_n(1 - gs) + cs]. \tag{20}$$

The essence of the SW proposition is contained in (19). If agents monitor s then the behavior of the quantity q is independent of the reaction coefficient g. On the other hand, (20) shows that q is affected by g if no monitoring occurs.

From (10), (19), and (20) the mean and variance of output per person, as seen at the beginning of stage 3 (i.e., before the exact value of s has been ascertained), can be expressed as

$$E(q_n^1) = E(q_n^1 \mid s) = E(q_n^2) = \bar{q}_n \neq E(q_n^2 \mid s), \tag{21}$$

$$\text{var } q_n^1 = \sigma_b^2\bar{m}_n^2 + \sigma_s^2(\beta^2 + \sigma_b^2)\sigma_c^2, \tag{22}$$

$$\text{var } q_n^2 = \sigma_b^2\bar{m}_n^2 + \sigma_s^2(\beta^2 + \sigma_b^2)[\sigma_c^2 + (\gamma - g\bar{m}_n)^2]. \tag{23}$$

From (6) and (17)–(21) the utility that each agent expects at the end of stage 2 may be expressed as

$$V_n^i(g) = EV(M/P_n^i, q_n^i, x, 1) = \bar{m}_n + \bar{q}_n - (k/2)\bar{q}_n^2 - (k/2)\,\text{var}\, q_n^i;\ i = 1, 2. \tag{24}$$

It follows from (22)–(24) that the gain to society from everyone monitoring the indicator (i.e., $V_n^1[g] - V_n^2[g]$) arises solely from the consequent reduction in the variance of output. This reduction in variance yields a welfare gain because the marginal utility of more production is (by assumption) a constant but the marginal cost is increasing, so that total utility is a concave function of the level of output.

[11] An almost identical condition was *assumed* to hold by Fischer (1977, pp. 195–97) and by Phelps and Taylor (1977, pp. 166–70).

III. The Monitoring Decision

Equations (17)–(24) show how each agent's price, output, and expected utility will depend on the monitoring decision made in stage 2 and the reaction coefficient chosen in stage 1. This section analyzes the monitoring decision.

Let z_m (> 0) denote the cost, in utils, of monitoring the indicator. This cost will generally equal $z_c + z_p$ where z_c is the cost of collecting the information (i.e., of determining the exact value of s) and z_p is the cost of processing the information (i.e., of computing and posting the optimal price given the value of s). Each agent will choose to monitor if the gain to him in expected utility is at least as great as z_m.

The size of this gain will depend on whether or not the other agents are monitoring. Consider first the case in which none of the other agents is monitoring. In this case the agent will derive the expected utility $V_n^2(g)$, as determined by (24), if he too chooses not to monitor. But if he chooses to monitor, the market price will continue to be P_n^2, as in (18), and he can now profit from knowing the exact value of the indicator before setting his own price. Because he can infer, as before, the exact value of P before determining his relative price, that relative price will be the one that satisfies the first-order condition (12) with $I = \{s\}$ and $P = P_n^2(\bar{M}, g, s)$. That is,

$$r_n^1(P_n^2, \bar{M}, g, s) = 1 + s\beta(\gamma - g\bar{m}_n)/(\bar{q}_n + n). \tag{25}$$

Let $V_n^3(g)$ denote the expected utility of the agent who monitors when no one else is monitoring. This can be obtained by substituting from (25) into (6) and using $q = q_n^3 \equiv q_n^2 + n(1 - r)$ with q_n^2 given by (20). As a result of this calculation it can be seen that the gain in expected utility from deciding to monitor when others are not monitoring is

$$V_n^3(g) - V_n^2(g) = \phi(n)\sigma_s^2\beta^2(\gamma - g\bar{m}_n)^2 \tag{26}$$

where $\phi(n) = (1 + nk)^2/2n(2 + nk)$.[12] Note that

$$\phi(n) > (k/2) \text{ and } \phi(n) \to (k/2) \text{ as } n \to \infty. \tag{27}$$

The nature of this gain is easy to see. The expected output of the agent is not affected by the decision to monitor but its variance is, and, as we have argued, a reduction in variance increases expected utility. Specifically, whenever the level of aggregate demand is high ($s > 0$) the monitoring agent will, according to (25), raise his price, thereby damping the increase in his sales (assuming that $g < \gamma\bar{m}_n^{-1}$). Likewise, when $s < 0$ the monitoring agent will bolster his sales by reducing r. Not only does this reduce variability it also adds to the agent's ex-

[12] This result is also derived in the Appendix.

pected revenue. In the limiting case of perfect competition, (25) implies that these changes in r will be imperceptible. In this case the agent is able to adjust the mean position of his random sales schedule whenever a nonzero value of s is observed so that the expected value of his sales *conditional upon* s is equal to $\bar{q}$,[13] and the entire gain to the agent from deciding to monitor arises from the resulting reduction in the variance of sales. In particular, (26) and (27) imply that the gain is equal to $(k/2)$ times this reduction in variance.[14]

Next, consider the case in which all of the other firms are monitoring. In this case the agent will derive the expected utility $V_n^1(g)$ as determined by (24) if he too chooses to monitor. But if he chooses not to monitor then the price will continue to be P_n^1 as in (17), and the agent will lose by not being able to foresee, before setting his own price, the variations in s and in P_n^1 that his competitors are taking into account. The nonmonitoring agent will choose a price $\hat{P}$ so as to maximize $EV[M/P_n^1, q_n^1 + n(1 - \hat{P}/P_n^1), x, \hat{P}/P_n^1]$. Let $V_n^4(g)$ denote the expected utility of the agent who does not monitor when others are monitoring. Then, as the Appendix demonstrates, the gain in expected utility from deciding to monitor when others are monitoring is

$$V_n^1(g) - V_n^4(g) = \psi(n) \frac{\operatorname{var} e_n}{1 + \operatorname{var} e_n} \tag{28}$$

where $e_n = P_n^1(\overline{M}; g, s)^{-1}[EP_n^1(\overline{M}; g, s)^{-1}]^{-1}$ is the reciprocal of the market price, expressed as a ratio to its mean value, and where $\psi(n) = n + n^2k/2$. Note that

$$\psi(n) \to \infty \text{ as } n \to \infty. \tag{29}$$

If everyone is to monitor then the gain from deciding to monitor when everyone else is monitoring must be at least as great as the monitoring cost; that is,

$$V_n^1(g) - V_n^4(g) \geqq z_m. \tag{30}$$

Likewise, if no one is to monitor, it is necessary that

$$V_n^3(g) - V_n^2(g) \leqq z_m. \tag{31}$$

From (26) and (28) it follows that for large enough n at least one of these conditions must be satisfied. But it also follows that both of them

[13] The limiting case of perfect competition is denoted by the absence of the subscript n on the variables $\bar{q}_n$, $\bar{m}_n$, V_n^1, $\hat{g}_n$, etc.

[14] Notice that we are not allowing a monitoring agent to gain by selling his information. Allowing a market for information would complicate the analysis without affecting our main result. For in the case that we focus on below where an activist policy dominates all others the private value of information is less than the cost of processing it. In this case no one would willingly pay the monitoring agent anything for his information.

may be satisfied. In other words, convention appears to play an important role in the monitoring decision. It may be that everyone will monitor if they all expect everyone else to, but no one will monitor if they all expect no one else to.[15] I shall assume that monitoring will occur if and only if (31) is violated—in other words, that the convention of monitoring will develop only when the convention of non-monitoring is unsustainable.[16] Notice, however, that when the reaction coefficient assumes the value

$$\hat{g}_n = \gamma/\bar{m}_n \tag{32}$$

then the convention of monitoring could never be sustained as long as $z_m > 0$, because from (17), (26), and (28),

$$V_n^1(\hat{g}_n) - V_n^4(\hat{g}_n) = V_n^3(\hat{g}_n) - V_n^2(\hat{g}_n) = 0. \tag{33}$$

In order to simplify the analysis further, let us assume the limiting case of perfect competition. Then, from (2), (26), (27), (31), and (32) the necessary and sufficient condition for monitoring to occur is

$$(k/2)\lambda_c\sigma_x^2\beta^2(1 - g/\hat{g})^2 > z_m \tag{34}$$

where $\lambda_c = \gamma^2/(\gamma^2 + \sigma_c^2)$ is a measure of the accuracy of the indicator, being the square of the coefficient of correlation between x and the conditional forecast γs of x based on the indicator, $\sigma_x^2 = \text{var } x$ is the variance in aggregate demand, and $\hat{g} = \gamma\beta k$, the limiting value of (32).

According to (34), the likelihood of monitoring occurring is larger (i) the larger is the slope, k, of the typical producer's marginal cost schedule, because the larger this slope the more the firm loses from variability in its demand that could be avoided by monitoring; (ii) the

[15] In such cases the private value of the information contained in s is larger when everyone else has the information than when no one else has it. This result may appear somewhat counterintuitive, but it arises from the following considerations. In stage 4 the agent who knows s may gain because this allows him to vary his price in anticipation of variations in (*a*) the variable x which is correlated with s, and (*b*) his competitors' price, which may be predicted once s is known. When n is large, then *b* will be the dominant factor if others are observing s, because variations in his competitors' absolute selling price imply variations in his own relative selling price, and thus large variations in his sales, if he is unable to anticipate them. But if no one else is observing s then *b* no longer applies because he does not need to observe s in order to anticipate his competitors' price. More generally, the result arises because the main use of the information is to allow the agent to coordinate his activities (specifically the activity of setting his nominal selling price) with the activities of other agents. As an analogy, the value of knowing that you should drive on the left in England would be much less if none of the other drivers knew it. This analogy also illustrates why multiple equilibria that must be selected by convention can arise under such circumstances.

[16] Thus we may say that an equilibrium with respect to any $(g, \bar{M}, n)$ consists of a monitoring decision $i \in \{1, 2\}$ and a price function $P_n^i(\bar{M}, g, s)$ such that (*a*) for any s the optimal relative price $r_n^i(P, \bar{M}, g, s)$ equals unity when $P = P_n^i(\bar{M}, g, s)$, and (*b*) the monitoring decision is $i = 1$ only if $V_n^1(g) - V_n^4(g) \geqq z_m$, and $i = 2$ if and only if $V_n^3(g) - V_n^2(g) \leqq z_m$.

more accurate is the indicator, because an inaccurate indicator conveys little useful information; (iii) the more variable is the level of aggregate demand, because there is little gain in monitoring a variable that is relatively constant; (iv) the larger is β, because there is little gain to monitoring a variable if its expected effect on demand is small; and (v) the further is the reaction coefficient from $\hat{g}$, because when $g = \hat{g}$ then, as can be inferred from comparing (22) with (23), the monetary authority is already accomplishing everything that private agents could accomplish by monitoring.

IV. Optimal Monetary Policy and the Incentive to Monitor

Three implications of this analysis are worth noting here. First, the SW proposition is valid for all values of g satisfying (34). For in this case monitoring will occur, and, as noted before, the behavior of output is unaffected by the value of g. According to (22) and (24) each agent's expected utility will also be unaffected by the value of g.

Second, whatever can be accomplished by monitoring can be accomplished just as well by having the monetary authority set $g = \hat{g}$. According to (34) no monitoring would occur, but according to (19), (20), and (24) the behavior of output and expected utility would be the same as if monitoring was occurring.

Third, as was argued in the introduction above, private agents may not have much incentive to monitor s even when the social gains from everyone deciding to monitor are large. That is, the private value of monitoring, $V^3(g) - V^2(g)$, may be small when the social value, $V^1(g) - V^2(g)$, is large. In particular, it follows from (22)–(24) and (26) that

$$[V^3(g) - V^2(g)] = \lambda_b[V^1(g) - V^2(g)], \tag{35}$$

where $\lambda_b = \beta^2/(\beta^2 + \sigma_b^2)$ is a measure of the predictability of the effect upon effective demand of changes in x, being the square of the coefficient of correlation between the actual effect bx and the predicted effect βx. If the value of the coefficient b is known with certainty, then the private and social values coincide. In the limiting case of pure uncertainty (when $[\beta/\sigma_b]^2 \to 0$) the private value is infinitesimal in comparison with the social value.

The reason for this divergence of private from social value is seen most easily in the case of a perfectly reliable indicator ($\sigma_c^2 = 0$) and a neutral monetary policy ($g = 0$). If everyone monitors then the price level will adjust to fluctuations in x so that the sum $m + x$ is constant and equal to $\bar{m}$, with output equal to $q^1 = b\bar{m}$. Thus everyone monitoring together will completely eliminate the effects of x upon output. An isolated agent monitoring when no one else is cannot stop $m + x = \bar{m} + x$ from fluctuating. He can, and will, through impercep-

tible changes in r alter the mean position of his demand curve so as to offset the predicted effect βx, but there will still remain an unpredicted effect $(b - \beta)x$. Specifically, by substituting (25) into (9) we see that his output will be $q^3 = b\bar{m} + (b - \beta)x$. Thus the social gain to monitoring comes from eliminating the total effect bx of changes in aggregate demand, whereas the private gain comes from eliminating only the predicted effect βx. The ratio λ_b of private to social gain is just the ratio of variances of the predicted and total effects.

Consider three different monetary policies. (i) Policy N is the neutral policy, $g = 0$, which is assumed to result in no administrative cost. (ii) Policy D is the disseminating policy suggested by Barro of setting $g = 0$ but also publicizing the value of s, which is assumed to result in the administrative cost z_d per person. (iii) Policy A is the activist policy of setting $g = \hat{g}$, resulting in the administrative cost z_a. Assume that the cost z_a is incurred whenever *any* nonzero value of g is chosen. Then policy A will dominate any other policy with $g \neq 0$, and we may restrict our attention to policies A, D, and N. Which policy is optimal will depend upon the relationships between the costs and benefits of monitoring and the administrative costs of the different policies. Let

$$V^s \equiv V^1(0) - V^2(0) = (k/2)\lambda_c \operatorname{var}(bx) \tag{36}$$

denote the social gain to monitoring and $V^p \equiv V^3(0) - V^2(0) = \lambda_b V^s$ denote the private gain to monitoring, when $g = 0$. Then according to (34) monitoring will never occur under policy A and will not occur under policy N if $\lambda_b V^s \leqq z_c + z_p$. Policy D is assumed to save the private agents any collecting costs, but if they are to use their knowledge of the value of s they must still incur the processing cost. Thus monitoring will not occur under policy D if $\lambda_b V^s \leqq z_p$. Table 1 depicts the cost per person of the different policies in the three situations indicated, where the cost is measured as $V^1(0)$ minus the expected utility resulting from the policy plus the monitoring and administrative costs. In each situation the optimal policy is the one with the least cost.

TABLE 1

COST PER PERSON OF MONETARY POLICY

	Policy		
Situation	N	D	A
I. $z_c + z_p < \lambda_b V^s$	$z_c + z_p$	$z_d + z_p$	z_a
II. $z_p < \lambda_b V^s \leqq z_c + z_p$	V^s	$z_d + z_p$	z_a
III. $\lambda_b V^s \leqq z_p$	V^s	$z_d + V^s$	z_a

NOTE.—N = neutral policy, D = disseminating policy, A = activist policy

Note that if the indicator is unreliable enough (i.e., λ_r small enough) or if effective demand is stable enough (var $[bx]$ small enough), *ceteris paribus*, then a neutral policy will be optimal because both of these factors make V^s small. Both of these factors also figure prominently in Milton Friedman's (1953, 1968) case against activist policy. For the same reason, a small slope to marginal cost (small k) favors a neutral policy. Note also that in situation III the disseminating policy is dominated by the neutral policy.

For our purposes the most important implication to note about table 1 is that an activist policy will be uniquely optimal whenever $z_a < V^s \leqq z_p/\lambda_b$; that is, if (i) the indicator is reliable enough, effective demand is unstable enough, or marginal cost rises fast enough that V^s exceeds the cost of administering an activist policy; and (ii) the cost of processing the information contained in the indicator is large enough or the predictability of the effect of changes in the level of aggregate demand is small enough that $V^s \leqq z_p/\lambda_b$. These conditions do not require the monetary authority to possess any cost advantage. If λ_b is very small, they may hold even if the cost of administering an activist policy is much larger than the private cost of monitoring. If the conditions hold then an activist policy is optimal because despite the large social gain to monitoring no one will do it even if the cost of collecting information is saved. For even under a disseminating policy the user of the information contained in the indicator must incur a processing cost, and, as we have seen, he will be led to underestimate the gains from using the information.

V. A Variant of the Basic Model: Monetary Policy and the Efficiency of Monitoring[17]

One special feature of the basic model employed above is that the terms m and x enter into the aggregate demand function (5) with the same coefficient. This is why the optimal reaction coefficient $\hat{g}$ always makes the expected level of output (conditional on s) equal to its optimal value, $\bar{q}$ (see [20]). When m and x have different random coefficients then Brainard's (1967) analysis suggests that the optimal value of the reaction coefficient will generally be smaller than this. Such a case is easily constructed by replacing the utility function (1) with the variant

$$U = K(\hat{q} - x)^{a}\hat{m}^{1-a} - (k/2)q^2 + 2x, \tag{37}$$

in which case the indirect utility function is still given by (6), but the

[17] The rationalization for activist policy presented in this section is similar to the one presented by Fane (1977).

effective demand function changes from (7) to

$$q = bm + x. \tag{38}$$

As in (19) and (20), the value of output with or without monitoring can be expressed as

$$q_n^1 = b(\bar{m}_n - \gamma s/\beta) + cs \tag{39}$$

$$q_n^2 = b\bar{m}_n(1 - gs) + cs. \tag{40}$$

Once again the expected value of output is

$$E(q_n^1) = E(q_n^1 \mid s) = E(q_n^2) = \bar{q}_n \tag{41}$$

and the variance of output is

$$\operatorname{var} q_n^1 = \sigma_b^2\bar{m}_n^2 + \sigma_s^2[(\sigma_c^2 + \gamma^2\sigma_b^2)/\beta^2] \tag{42}$$

$$\operatorname{var} q_n^2 = \sigma_b^2\bar{m}_n^2 + \sigma_s^2[\sigma_c^2 + (\gamma - g\bar{m}_n\beta)^2 + (g\bar{m}_n)^2\sigma_b^2]. \tag{43}$$

Let us again restrict our attention to the limiting case of perfect competition. As before, the optimal value g' of the reaction coefficient will be the one that minimizes var (q^2). That is,

$$g' = \lambda_b\gamma/\bar{q}. \tag{44}$$

If the activist policy $g = g'$ is pursued and no monitoring occurs, then

$$E(q^2 \mid s) = \bar{q} + (1 - \lambda_b)\gamma s. \tag{45}$$

As in Brainard's analysis, only the fraction λ_b of the predicted effect of a change in the indicator is offset by optimal monetary policy. In contrast to this, (41) implies that all of the predicted effect will be offset by the outcome of monitoring. Thus optimal monetary policy without monitoring can in this case accomplish more than monitoring. With private monitoring the real money supply will overreact to information about aggregate disturbances. Specifically, it follows that the social gain to monitoring when $g = g'$ is

$$V^1(g') - V^2(g') = (k/2)\lambda_c\sigma_x^2\sigma_b^2[1/(\beta^2 + \sigma_b^2) - 1/\beta^2] < 0. \tag{46}$$

Optimal monetary policy can be characterized in this variant model as we did in the basic model with results that are similar but with a complication, because even when $g = g'$ and the social gain to monitoring is negative, the private gain in this case is actually positive. Specifically,

$$V^3(g) - V^2(g) = (k/2)\lambda_c\sigma_x^2(1 - \lambda_b g/g')^2.^{18} \tag{47}$$

[18] See Appendix.

from which it follows that $V^3(g') - V^2(g') > 0$. Thus optimal monetary policy may require g to be somewhat larger than g' so that people will not be motivated to monitor.

For our purposes the important feature of optimal monetary policy is that the activist policy ($g = g'$) will be optimal whenever

$$(k/2)\lambda_c\sigma_x^2(1 - \lambda_b)^2 \leqq z_p + z_c < (k/2)\lambda_c\sigma_x^2 \tag{48}$$

and

$$z_a \leqq (k/2)\lambda_c\sigma_x^2\sigma_b^2[1/\beta^2 - 1/(\beta^2 + \sigma_b^2)]. \tag{49}$$

If (48) is satisfied then monitoring will occur under a neutral or disseminating policy but not under an activist policy. If (49) also holds then the activist policy is optimal because its cost (relative to $V^2[g']$) is z_a, whereas, from (46) and the fact that $V^1(g)$ is independent of g, the cost of each of the other policies will exceed the RHS of (49) by the sum of monitoring and administrative costs. As in the previous section, the monetary authority needs no cost advantage for these conditions to be satisfied if $\lambda_b < 1$. For example, they may be satisfied even if $z_a > z_c + z_p$, because the first term in (48) is just λ_b times the RHS of (49).

Thus, an activist monetary policy can be justified because the monetary authority, by avoiding the overreaction to the indicator that would occur with monitoring, is able to accomplish more than could be accomplished by private monitoring. The overreaction of private agents monitoring s occurs because each individual sees himself as being able to offset the effect of s by changing his relative price, the coefficient of which, n, is known with certainty. Thus, he will try to offset completely the predicted effects of a change in s. As Brainard has shown, this is the optimal policy when the effect of the price setter's action can indeed be predicted with certainty. But the effect of every agent trying to alter his relative price is a change in the absolute price level, the effect of which cannot be predicted with certainty. As Brainard has also shown, the optimal reaction when the effect of policy is uncertain is to attempt to offset less than 100 percent of the predicted effect of the disturbance. The crucial assumption in this argument is that the degree of uncertainty attached to the effects of relative-price changes is different from that attached to absolute-price changes. Private agents take the former into account but not the latter. In the particular model used above the former uncertainty is less than the latter, but this is not essential to the argument. In the reverse case private agents would *under*react to monitored aggregate disturbances, once again giving a rationale to an activist policy.

VI. Conclusion

In summary, this paper has argued that an activist monetary policy may be justified even when people's expectations are formed rationally if there are costs of gathering and processing information and if there is uncertainty concerning the exact values of parameters in the economic system. The justification that we have presented is twofold: first, that private agents may not be motivated to use information that can be used profitably by the monetary authority and, second, that even if they are motivated to use it they may not make as good use of it as could the monetary authority.

The formal argument has been presented with a simple model employing specific functional forms chosen for their tractability. Therefore, much work remains before the degree of generality of the results can be ascertained.

The first rationale that we provided for an activist policy should be distinguished from that of Fischer (1977) and Phelps-Taylor (1977). Their models are similar to ours in that prices are set in advance of trading rather than being determined by an independent auctioneer according to market-clearing conditions. However, their models allow activist monetary policies to improve welfare because they postulate a temporary rigidity in prices that prevents private agents from realizing all mutually advantageous and perceived gains from trade. This raises the question of whether the contractual arrangements implicit in such models would be competitively sustainable (see, e.g., Barro 1977). Our argument has gone one step further and shown conditions under which these contracts are in fact sustainable. We have not postulated any rigidity or imperfection that allows the monetary authority to react more quickly or with more flexibility to variation in the indicator than private agents can. Instead we have shown that private agents may *choose* not to act on as much information as the monetary authority possesses, or even as much as the monetary authority makes freely available to them, even when it would be Pareto improving for them to do so. In other words, this argument also provides an explanation for the fact that private contracts are typically written in nominal terms without being indexed to the money supply or other aggregate variables.

The paper has not addressed the question of whether or not the particular conditions derived above for the optimality of an activist policy actually exist in any real-world situation. The purpose has been mainly to argue that as a matter of principle the assumption of rational expectations does not imply the uselessness of an activist policy. It depends at least on the factors discussed above. Any bal-

anced evaluation of alternative monetary policies should also take into account at least two other important considerations. First, as Hayek (1945) has argued, the most important problem to be resolved is probably not what policy to pursue but whom to entrust with the task. An activist policy will work only if the authority can be entrusted or persuaded to pursue the public interest instead of his own and to pursue it competently. Second, as Hayek has also argued, the centralization of decision making is most easily rationalized when the decisions require the use of general knowledge rather than local, specific knowledge. In the model of the present paper every individual agent is essentially identical, so there is no clear distinction between these two kinds of knowledge. But more generally one would expect the aggregate disturbances to which monetary policy reacts to come under the heading of general knowledge, about which most individual agents can be expected to have little expertise. Thus a further argument in favor of an activist policy is that it allows society to centralize, and thus to bring to bear an efficient degree of expertise on, decisions concerning such general knowledge.

Appendix

This Appendix sketches a formal derivation of the private value of monitoring. In general terms, suppose that a decision maker has an information set I and his problem is to choose the value of decision variable x so as to

$$\max_{(x)} E(\alpha_0 + \alpha_1 x + \alpha_2 x^2 \mid I) \tag{A1}$$

where the α_i's are random parameters. Then the expected value of receiving the information contained in the new information set I' is the expected difference in the optimized value of (A1) resulting from replacing I by I'; that is,

$$E\{[E(\alpha_1 \mid I)]^2/4E(\alpha_2 \mid I) - [E(\alpha_1 \mid I')]^2/4E(\alpha_2 \mid I')\}. \tag{A2}$$

1. To derive (26) from (A2) note that in this case $\alpha_1 = (q_n^2 + n)(1 + nk)$, $\alpha_2 = -(n/2)(2 + nk)$, $I = \emptyset$, and $I' = \{s\}$.

2. To derive (28), let x be the *average* relative price, $PE[P_n^1(\bar{M}, g, s)^{-1}]$, and note that in this case $\alpha_1 = e_n(q_n^1 + n)(1 + nk)$, $\alpha_2 = -e_n^2(n/2)(2 + nk)$, $I = \emptyset$, and $I' = \{s\}$.

3. To derive (48) proceed exactly as in deriving (26) and take the limit as $n \to \infty$.

References

Barro, Robert J. "Rational Expectations and the Role of Monetary Policy." *J. Monetary Econ.* 2 (January 1976): 1–32.

———. "Long-Term Contracting, Sticky Prices, and Monetary Policy." *J. Monetary Econ.* 3 (July 1977): 305–16.

Brainard, William C. "Uncertainty and the Effectiveness of Policy." *A.E.R. Papers and Proc.* 57 (May 1967): 411–25.

Fane, G. "Stabilization Policy in Models with Rational Expectations and Uncertainty." Mimeographed. Australian National University, 1977.

Feige, Edgar L., and Pearce, Douglas K. "Economically Rational Expectations: Are Innovations in the Rate of Inflation Independent of Innovations in Measures of Monetary and Fiscal Policy?" *J.P.E.* 84, no. 3 (June 1976): 499–522.

Fischer, Stanley. "Long-Term Contracts, Rational Expectations, and the Optimal Money Supply Rule." *J.P.E.* 85, no. 1 (February 1977): 191–206.

Frenkel, Jacob A. "Inflation and the Formation of Expectations." *J. Monetary Econ.* 1 (October 1975): 403–21.

Friedman, Benjamin M. "Optimal Expectations and the Extreme Information Assumptions of 'Rational Expectations' Macromodels." *J. Monetary Econ.* 5 (January 1979): 23–41.

Friedman, Milton. "The Effects of a Full-Employment Policy on Economic Stability: A Formal Analysis." In *Essays in Positive Economics*. Chicago: Univ. Chicago Press, 1953.

———. "The Role of Monetary Policy." *A.E.R.* 58 (March 1968): 1–17.

Grossman, Sanford J., and Stiglitz, Joseph E. "Information and Competitive Price Systems." *A.E.R. Papers and Proc.* 66 (May 1976): 246–53.

Hayek, Friedrich A. von. "The Use of Knowledge in Society." *A.E.R.* 35 (September 1945): 519–30.

Hirshleifer, Jack. "The Private and Social Value of Information and the Reward to Inventive Activity." *A.E.R.* 61 (September 1971): 561–74.

Laidler, David E. W. "Money and Money Income: An Essay on the 'Transmission Mechanism.'" *J. Monetary Econ.* 4 (April 1978): 151–91.

Lucas, Robert E., Jr. "Expectations and the Neutrality of Money." *J. Econ. Theory* 4 (April 1972): 103–24.

———. "An Equilibrium Model of the Business Cycle." *J.P.E.* 83, no. 6 (December 1975): 1113–44.

Phelps, Edmund S., and Taylor, John B. "Stabilizing Powers of Monetary Policy under Rational Expectations." *J.P.E.* 85, no. 1 (February 1977): 163–90.

Poole, William. "Rational Expectations in the Macro Model." *Brookings Papers Econ. Activity*, no. 2 (1976), pp. 463–505.

Rutledge, John. *A Monetarist Model of Inflationary Expectations*. Toronto: Lexington, 1974.

Sargent, Thomas J. "A Classical Macroeconometric Model for the United States." *J.P.E.* 84, no. 2 (April 1976): 207–37.

Sargent, Thomas J., and Wallace, Neil. "'Rational' Expectations, the Optimal Monetary Instrument, and the Optimal Money Supply Rule." *J.P.E.* 83, no. 2 (April 1975): 241–54.

Shiller, Robert J. "Rational Expectations and the Dynamic Structure of Macroeconomic Models: A Critical Review." *J. Monetary Econ* 4 (January 1978): 1–44.

Taylor, John B. "Monetary Policy during a Transition to Rational Expectations." *J.P.E.* 83, no. 5 (October 1975): 1009–21.

Equalizing Discrimination and Cartel Pricing in Transport Rate Regulation

Kenneth D. Boyer

Michigan State University

There are two possible outcomes of transport regulation: (1) maintaining a carrier cartel and (2) imposing equalizing discrimination against advantaged and in favor of disadvantaged shippers. Both functions have required a complex rate structure to enforce the respective forms of price discrimination. Using a sample of freight bills from motor carriers and railroads, this paper demonstrates that the principal result of motor carrier regulation has been to maintain a cartel of truckers, while railroad regulation has thwarted the wishes of the railroad cartel by imposing equalizing discrimination on weak and strong shippers. Motor carrier rates respond in an economically rational manner to costs and shipper bargaining power; rail rates are either unresponsive or perversely responsive to the same factors. Deregulation should have divergent effects in the two industries.

Deregulation is a policy that has been recommended in recent years for some but not all regulated industries (see, e.g., Council of Economic Advisers 1979). Candidates for deregulation are most often in the transport sector and never in the traditional public utilities such as electricity and natural gas distribution. Why is transport regulation consistently singled out for attention? Also curious are the extraordinary attempts of motor carriers to retain the regulatory fetters, while the railroads make somewhat half-hearted attempts to remove theirs (see, e.g., Williams 1978). What is it about the type of regulation that makes

This paper was improved by advice and comments from W. G. Shepherd, A. S. Lang, H. Trebing, G. Wilson, P. Schmidt, R. Rasche, D. Suits, S. Tainter, and an anonymous referee.

[*Journal of Political Economy*, 1981, vol. 89, no. 2]
© 1981 by The University of Chicago. 0022-3808/81/8902-0001$01.50

one transport industry wish to retain—at all costs—what another transport industry is willing to shed?

As this paper explains, the answer to both questions is found in the institution of transport rate making. There are two key differences between traditional forms of price regulation and transport rate controls. First, in most instances, regulation involves category pricing, while transport rates are tailored to individual buyers. Second, transport rates are typically the result of collective rate making, while most other regulated prices are not. Together, these aspects allow the regulatory authority to bestow advantages and disadvantages on individual communities and industries; they also allow the industry to stabilize a sellers' cartel.

This paper demonstrates that there would be two distinct results from deregulating the motor carrier and railroad industries since there are divergent outcomes of regulation in the two modes. Current motor carrier rates show classic signs of cartel rate making; railroad rates reveal Interstate Commerce Commission (ICC) attempts to equalize advantages and disadvantages of shippers and localities. Deregulating the trucking industry would inhibit cartel rate making; deregulating the railroad industry would lead to a more economically rational pattern of prices, with some rates rising and some falling.

To emphasize the unusual nature of transport pricing, Section I shows what category pricing would be like in transport. Section II shows how individualized rates and collective rate making lead to unpredictable outcomes in the regulatory process. Unlike category pricing, transport rate making is ideally suited for either cartel pricing or enforcing equalizing discrimination. Section III investigates the current motor carrier and railroad rate structures and finds that different principles are apparent in the pricing of the two modes. The paper concludes with implications of the findings for deregulation arguments.

I. Transport Prices Are Individual Rather than Category Rates

Compared with motor carriers and railroads, the rate structure for most regulated industries is the model of simplicity: The customer's class of service is determined, and the block pricing for the class is applied. Such category pricing has two important implications. First, a customer's charges can be raised or lowered only by altering the prices assessed everyone in that category. Customers cannot demand special price concessions and are protected from unusually high rates by the presence of other members in the category. Category pricing prevents bargaining over individual charges between suppliers and customers.

The second implication is for regulators. Because the pricing is

relatively simple, the regulatory authorities can make informed decisions about each price. The costs of providing service to a category of customers can, at least in principle, be calculated with some accuracy; again in principle, the regulators can check the supplier's cost calculations with a limited staff.

The costs of providing transport service are quite variable (Harris 1977; Spady 1979). Thus category prices for railroads and motor carriers would be more complicated than the price structures of most regulated industries. Such a scheme would, nonetheless, be feasible. For example, railroad car rentals (which could change daily according to availability) could be separated from haulage charges. Haulage charges could be divided into categories depending on the number of cars in the train, the regularity of movement, and so forth. Track rentals could also be charged and graduated according to expected maintenance charges on a link traversed. Congestion tolls could be assessed if a shipment used a link with excessive traffic. Separate charges could be assessed for using switching locomotives and renting yard space.

Similarly, motor carriers could be compelled to charge separate prices for separate services: one price for pickup, another for terminal sorting, another for use of line-haul service on a certain link, and so on. Total transport charges would then be determined by summing the costs of the service categories used.

As long as a transporter could charge for only a limited number of service categories, this scheme would mean that a customer's charges could be altered only by changing the rates of all those in a category, and that, by limiting the number of prices reviewed by the ICC, a competent staff could determine the relation between the price and the cost of a service. As demonstrated below, if transport regulation had the nature of category pricing, there would be scant reason to expect intense political pressure either to preserve or to remove the regulatory apparatus.

II. The Implications of Pricing Services Individually

Most transport sales are not based on category prices but use a scheme of exception or commodity rates. Each price is tailored to the individual buyer (Locklin 1972). As demonstrated below, this has inevitable consequences for the relations among carrier, customer, and regulator.

A system of class rates for motor carriers and for railroads does exist. Under this scheme, each commodity is given a class rating and each origin-destination pair is given a rate base number. The transport price is the product of these two numbers. The class-rate tariff is

thus a two-category scheme. It makes pricing simple and thus capable of intelligent oversight by the ICC; it also means rates cannot be changed for one shipper without simultaneously changing them for all others in the same class.

Because the class-rate scheme is only a two-category system, it cannot adequately reflect carrier costs which vary in more than two dimensions. In particular, neither economies of traffic density nor the likelihood of receiving a backhaul is accounted for in the class-rate system.

These unusual cost circumstances can be accommodated through the commodity-rate and exception-rate schemes which allow the quotation of a special price for shipping a good defined at greater than seven-digit precision between any two stations. That rates can be quoted so precisely reveals that they apply in practice to individual shippers. Commodity and exception rates are the opposite of category rates—they are individual rates.

Commodity and exception schemes are now almost the universal pricing method for rail shipments, and they are the predominant form of truckload rates. The class-rate system is reserved for occasional shippers and for much less-than-truckload traffic.

To understand the nature of transport rate regulation is to understand the interaction between the ICC and those parties that propose a change in a commodity or exception rate. More than 85 percent of motor carrier rate changes are reductions proposed by carriers or, more frequently, the rate bureaus to which they belong.[1] The universal rationale is to retain or gain traffic that would otherwise be moved by another mode.

Under individual pricing, shippers may gain advantages over competitors if their own rates are lowered but other shippers' are not. To bring this about, a shipper must convince the carrier that either (1) existing traffic will be diverted without the lower rate or (2) new, formerly uneconomical markets will be opened up if the rate is lowered. A shipper has two hurdles: First, the shippers must convince the carriers who would offer the rate that the threat of traffic gain or loss is genuine;[2] second, should the rate be protested, the ICC must rule

[1] This figure was quoted by M. A. Godecker of the Central States Motor Freight Bureau My understanding of regulatory procedure in this paper benefited by discussions with Godecker and J. R. Ahlstrom of the Western Truck Line Committee. Substantial alterations in the procedures for regulating railroads and motor carriers were passed by Congress in the time subsequent to the writing of the paper. The forms of regulation discussed in this paper were those practiced in through the spring of 1980

[2] Motor carriers claim that collective rate making is necessary to protect them from shippers' opportunistic misrepresentation of the likelihood of traffic diversion.

on it. To go into effect the rate must be declared just, reasonable, and nondiscriminatory.

Precisely what criteria the ICC uses to judge rates have never been defined. The commission does demand cost evidence in the form of standard cost statements. It recognizes, however, that the costs summarized in the statements are average costs, and thus additional evidence of haulage costs is used; the standard piece of evidence is the rate of analogous hauls.

The basic ICC regulatory procedure assures a balance of rates on analogous shipments. A just, reasonable, and nondiscriminatory rate is one which is not out of line with rates on analogous hauls. What is in practice regulated, then, is the rate relationship among analogous or competing hauls.

The result of this regulatory procedure has been a transport rate structure of fantastic complexity: Hundreds of billions of individual prices have been established and approved. In contrast, under the typical category scheme used in most regulated industries, there are rarely more than 100 separate rate categories. The complexity of the transport rate structure precludes staff investigations of price-cost relationships for individual rates. Indeed, the very complexity of the rate structure discourages cost finding. With traffic equal to the sum of large numbers of unique movements, virtually all costs will be found to be joint or common and thus logically or practically untraceable.

This inability to cost individual movements has distinct advantages both to carriers and to the ICC since it increases the discretion of each party to pursue separate goals in the regulatory process. The lack of reliable cost figures frees the rate bureaus to engage in demand elasticity–based pricing which is implicit in the policy of granting rate reductions only to those shippers who can credibly threaten traffic diversion. Without cost data, discrimination is extraordinarily difficult to prove. The complaint of a firm that lower rates are charged to a rival can be answered by the apology that the rate concession was necessary to retain traffic. It is, of course, crucial to the proper working of the scheme that price-induced traffic diversions to firms with identical cost structures be inhibited, otherwise the process of collective rate making will evolve toward cost-based prices.

The difficulty of relating prices to costs gives wider latitude to the regulatory authorities to pursue what Hilton (1972) has called "the basic behavior of regulatory commissions." Hilton identified the goal of regulation as dissipation of monopoly profits through cross-subsidization; a preferable description in the case of transport would be requiring discrimination which equalizes the advantages and disadvantages of different purchasers of transport service. The phrase

"equalizing discrimination" is preferred to the more familiar title of "cross-subsidization" since the latter implies that the carrier uses profits from one service to finance another. But if carrier operations generally run a deficit, no service is paying for any other. The term "equalizing discrimination" emphasizes that the ICC is interested in equalizing the conditions of advantaged and disadvantaged shippers rather than in having one party pay the bills of another.

Equalizing discrimination is the natural result of attempting to prevent "unjust" or "extractive" discrimination (i.e., pricing to extract rents from shippers and localities) without using a cost comparison. For example, the volumes generated by a large shipper typically will produce both lower carriage costs and many eager carriers, while a small shipper will have fewer alternatives and higher carriage costs. To prevent exploitation of the small shipper's lower demand elasticity, in the absence of a reliable cost comparison, the regulatory authorities have no option other than to require equal prices for both shippers. The burden of this equalizing discrimination naturally falls on the advantaged shipper.

Explicit examples of equalizing discrimination can be found in policies that require identical prices for shipment of goods from the Midwest to any East Coast port regardless of distance, require low rates on California citrus fruit and high rates on Florida produce so that both can compete in Northeast markets, and discourage the quotation of unit train rates even when there are obvious cost advantages (MacAvoy and Sloss 1967).

The pressures on the ICC to practice such equalizing discrimination are most intense in the case of relatively low value, homogeneous commodities such as agricultural products, coal, steel, and paper. Demand for these goods is most sensitive to delivered prices, and their profitability is most strongly affected by transport prices. These are also the goods carried predominantly by rail. In contrast, the ICC gets less intense pressure to use transport prices to equalize the conditions of manufacturers of finished goods—the most important customers of motor carriers. This is because demand is less sensitive to transport prices and because transport rates are a smaller part of delivered cost.

Not only is there less pressure on the ICC to produce equalizing discrimination for those commodities hauled primarily by motor carriers, but the commission has less leeway, since carriers can use their own trucks if common-carrier rates get too high. One would expect, therefore, that the ICC would give more care to regulating rail than motor carrier rates: Their range of actions and the participants' interests are greater than in truck rates.

In the next section the rate structures of the two modes are investi-

gated to test the hypothesis that the regulatory influence of the ICC is more evident in railroad than in motor carrier pricing.

III. Empirical Evidence of Different Regulatory Regimes for Railroads and Motor Carriers

To demonstrate the dual outcome of ICC regulation, individual rates were regressed on factors which proxy demand elasticities and costs of providing service. A cartel unthwarted by the ICC will have rates that vary directly with the costs of service and inversely with demand elasticities.

It is a good deal more difficult to specify what would determine rates under equalizing discrimination. However, two aspects are likely: First, cost savings of a large shipping volume will not be passed along to customers in the form of lower charges; and second, having fewer shipping options is likely to lower rather than to raise rates since a low demand elasticity indicates a vulnerable shipper.

The complexity of rail and motor carrier rate structures derives from the discriminatory aims of both the rate bureaus and the regulators. But this same complexity precludes investigating rate structures in any manner other than explaining the patterns of actual charges for purchased services.[3] A sample of rail freight bills has long been available from the ICC, but the similar data file for motor carriers is the proprietary information of motor carrier rate bureaus.[4] This study was made possible by obtaining the latter file for the year 1972 and comparing it with the similar rail file.

The six analyzed commodities (the same for both rail and motor carriers) are perhaps not typical of either motor carrier or rail traffic as a whole; the two modes do not have identical traffic mixes.[5] Rather, the commodities were chosen to be representative of the goods contested by both railroads and motor carriers.

[3] The use of a sample of actual freight bills to model the rate structure rather than the alternative procedure of using rates printed in formal tariff statements is preferred for several reasons: (1) There are always numbers of rates available to any shipper, and it is not clear which would be the governing rate; (2) charges are sometimes different from those printed in tariffs; and (3) a large part of the rate structure is nominal in the sense that no freight is ever moved under these rates.

[4] The rail data are the ICC "One Percent Waybill Sample of Carload Terminations" (U.S. Interstate Commerce Commission 1972). The motor carrier data are from the "All-Bureau Continuing Traffic Study for 1972" (Rocky Mountain Motor Freight Bureau 1972).

[5] The six commodities are: prepared feed NEC, assigned STCC numbers 2042129 and 2042190 and NMFC number 72910; wrapping paper, STCC 2084120/21, NMFC 111510, 2621490, 151800; pulpboard or fiberboard, STCC 2631115/16120, NMFC 151310, corrugated fiberboard or pulpboard boxes, STCC 2651141, NMFC 29275; tires, STCC 3011110, NMFC 157230. For a complete description of commodity selection procedures and data filters, see Boyer (1978).

The charges for shipping a hundredweight of each of the six commodities between various end points were studied. One of the strongest influences was found to be the size of the shipment. In the Appendix, I describe the procedure used to adjust the reported charge per hundredweight in each freight bill for variations caused by differences in loading weight. The dependent variable whose variation is explained by the regressions in this section is a least-squares forecast of what the charge per hundredweight on a freight bill would have been had the shipment been made at a standard size of 340 cwt for trucks and 680 cwt for trains.[6] Only boxcar movements of rail commodities were considered.

The independent variables which could be generated from the freight bill samples were the length of the haul (*MI*); the number of shipments of this commodity, on this haul, in the sample (*SHP*); the number of alternative destinations open to a shipper of this commodity from this origin (*DST*); the number of origins for this commodity from which a receiver at this destination can choose (*ORN*); and existence of active competition from the other regulated mode of transport (*INT*). It can be assumed that costs per hundredweight increase with distance and decrease with shipping volume. Similarly, large values for *DST* and *ORN* as well as a positive value for *INT* should indicate a shipper with a higher demand elasticity. In addition, the existence of private trucking suggests that shipping volume will be positively related to demand elasticities.

The functional form for these regressions was:

$$\begin{aligned}\ln(\mathrm{RATE}_i) = {} & b_1(\ln MI_i) + b_2(\ln MI_i)^2 + b_3(\ln MI_i)^3 \\ & + b_4 M(\ln MI_i - \ln 100)^3 + b_5 N(\ln MI_i - \ln 500)^3 \\ & + C_1 \ln(SHP_i) + C_2 \ln(DST_i) + C_3 \ln(ORN_i) + C_4(INT) \\ & + \sum_{j=1}^{n} d_j D_{ji} + e_i,\end{aligned}$$

where

RATE = cents per hundredweight shipping charges (see Appendix);
MI = length of haul;
M = 1 if $MI_i > 100$, 0 otherwise;
N = 1 if $M_i > 1{,}000$, 0 otherwise;
SHP = number of freight bills in sample corresponding to this haul and commodity;

[6] These correspond to truckloads of 17 tons and carloads of 34 tons which were taken to be representative in Meyer et al. (1959).

DST = number of different destinations for freight bills in sample with this origin and commodity;

ORN = number of different origins for freight bills in sample having this destination and commodity;

INT = 1 if there is a freight bill in both samples with this commodity, destination, and origin and 0 if there is no intermodal competition; and

D_{ji} = 1 if observation is for commodity j and 0 if observation is not for commodity j.

The functional form contains a cubic spline of distances with knots placed at 200 and 1,000 miles.[7] I used this piecewise cubic approximation of the distance taper to avoid prior constraints on the shape of the distance relationship.

The regressions in table 1 show the results for the motor carrier rate structure. Table 1 also contains two rail regressions, one for the complete set of six commodities and one for a restricted set of the four goods for which haulage weight/rate relations could be estimated (see Appendix).

As reported in table 1, the results strongly support the conjecture that motor carriers use a cartel rate structure unhampered by government interference. Large shippers do get lower rates (the elasticity of rates with respect to the number of shipments is −.039) as predicted by the cartel model; the cost of serving large shippers is lower and the likelihood of switching to private trucking is higher. The coefficient on the number of destinations used by the seller (*DST*) and the number of origins used by the receiver of a shipment (*ORN*) similarly confirm that the greater the number of competing hauls, the lower the rates. Finally, the coefficient on *INT* suggests that motor carrier rates are approximately 6 percent lower where there is actual competition from railroads than where there is not. All of these findings are consistent with cartel rate making in the trucking industry.

The contrast to the rail results is striking. The response of rail rates to larger shipping volume (*SHP*) is insignificant. A large number of destinations (*DST*) lowers a shipper's rates, but a larger number of origins (*ORN*) does not similarly lower rates. And, finally, the presence of motor carrier competition (*INT*) is associated with higher, not lower, rail rates. With the exception of the negative sign on *DST*, the remaining results are those predicted by a policy of equalizing discrimination where rates are either unresponsive or perversely responsive to cost conditions and demand elasticities. The negative

[7] On estimating spline functions with ordinary least squares, see Suits, Mason, and Chan (1978)

TABLE 1

DETERMINANTS OF REGULATED FREIGHT RATES, CUBIC SPLINE REGRESSIONS

	Motor Carrier, Eq (1)		Rail			
			Eq. (2)		Eq. (3)	
ln *MI*	3.85	(3.71)	−17.35	(4.53)	−20.64	(4.87)
(ln *MI*)²	−.77	(.73)	3.16	(.89)	4.11	(.96)
(ln *MI*)³	.05	(.04)	−.22	(.058)	−.26	(.06)
(ln *MI* − ln 100)³	−.04	(.06)	.22	(.074)	.27	(.079)
(ln *MI* − ln 500)³	−1.92	(.10)	.23	(.10)	.27	(.11)
SHP	−.037	(.014)	−.011	(.012)	−.008	(.012)
DST	−.0062	(.0042)	−.014	(.006)	−.014	(.006)
ORN	−.012	(.1153)	.0018	(.0070)	−.00031	(.0071)
INT	−.062	(.021)	.060	(.020)	.064	(.021)
Dummy variable:						
3011110	.12	(.0079)	.46	(.015)	.48	(.015)
2651141	.26	(.016)	.51	(.028)	. . .	. . .
2631115	−.29	(.013)	.067	(.0090)	.09	(.008)
2621490	−.12	(.013)	.13	(.017)	.15	(.017)
2084120	−.12	(.037)	−.11	(.066)	. . .	. . .
2042190	−.24	(.016)	−.61	(.014)	−.58	(.014)
R^2	.54		.82		.83	
N	6,472		4,062		3,836	
F	658.78		904.06		855.56	

SOURCES — U.S. Interstate Commerce Commission 1972, Rocky Mountain Motor Freight Bureau 1972.
NOTE — Dependent variable = fitted cents/hundredweight for end-point pairs. Numbers in parentheses are SEs.

coefficient on *DST* indicates that there is some rational response to demand elasticities as well—an indication of some cartel rate making within a structure whose primary characterization is enforcing equalizing discrimination.

The regressions in table 1 give other evidence consistent with the hypothesis that railroads are thwarted by regulatory authorities while motor carriers set rates just as a successful trucker's cartel would wish. The first evidence comes from the spline variables on mileage. The mileage tapers, expressed on the vertical axis as a percentage of charges for a 500-mile haul, are graphed in figure 1.[8] The two modes appear to be very similar; the rail taper has unusually low short-haul rates and relatively high charges for the longest hauls. This is precisely the opposite of what would be generated by cost-based rates and contradicts the basis for many studies of optimal intermodal traffic allocation (Friedlaender 1969; Levin 1978). The slopes and positions of the two curves indicate that the observed propensity for long hauls

[8] On the safe assumption that rail rates will be lower than road rates at the standard 500-mile haul, the results suggest that rail and truck rates either diverge as mileage increases or that the rail curve cuts the motor carrier rate curve from below.

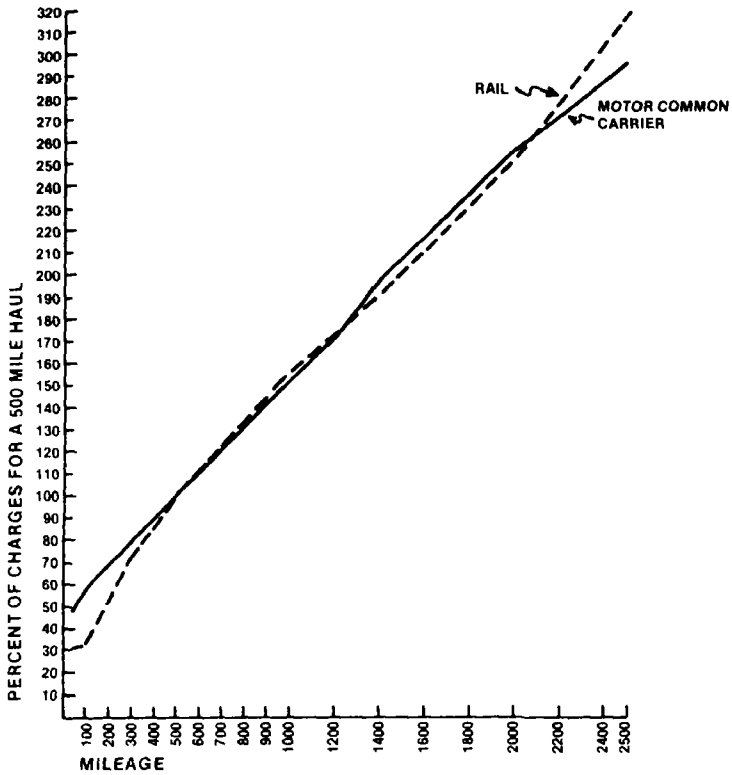


FIG. 1.—Mileage tapers for rail and motor carriers. Source: spline coefficients estimated in table 1.

to be shipped by rail and short hauls by truck is not the result of shippers responding to cost-based rates but, rather, must be due to a convergence of service qualities.

The results are consistent with an ICC policy of "fair sharing traffic" among modes. Rail rates are held high for long hauls to allow motor carrier participation and are allowed to fall for short hauls where motor carriers do not have to fear serious traffic diversions.[9]

A final piece of evidence for the differing regulation of motor carrier and railroad rates comes from the commodity-specific dummy variables in table 1. Converting logarithmic coefficients to percentages yields the data in table 2. The numbers are average rates for each of the six commodities. The variation in rail rates explainable by com-

[9] On fair sharing of traffic, see the "Ingot Molds Case" (American Commercial Lines, Inc., et al. v. Louisville and Nashville Railroad Co. et al. 1968). Note that the evidence is also consistent with one or the other, but not both, of the modes of pricing under cartel rate making.

TABLE 2

VALUE OF SERVICE PRICING

	Percentage of Charges on Average Commodity		
	Motor Carrier	Rail	
Commodity	from Eq. (1)	From Eq. (2)	From Eq. (3)
3011110	113.22	159.44	163 02
2651141	130.31	167.35	...
2631115	74.41	106.97	109.46
2621490	87.96	114.37	116.72
2084120	88.58	88.95	...
2042190	78.01	53 80	55.49
SD	21.82	42.90	44.05

SOURCE.—Table 1, antilog of commodity dummy variables.

modity value is more than twice as high as that for motor carrier rates: The column standard deviations are 42.9 percent for rail as opposed to 21.8 percent for truck. This suggests that rail rates are more closely related to commodity type than are truck charges.[10] This finding is consistent with a railroad industry which maintains an obsolete system of value-of-service pricing and a trucking industry which has adapted to an environment of intermodal competition.

IV. Conclusion

The evidence in this paper shows that both motor carrier and railroad rates are different from those that would be observed without regulation. The pattern of rates suggests, however, that regulation of the two modes has had different results. Perhaps the strongest support for the hypothesis that railroads and motor carriers are subject to different types of regulation comes from the public posture of the two modes toward deregulation. Within the last years, the Association of American Railroads has endorsed a "deregulation" proposal. However, the proposal does not suggest that collective rate making be abandoned. In contrast, the American Trucking Associations is vehement in its opposition to any changes in transport regulations. This is consistent with the findings of this paper: (1) Motor carrier regulation ratifies a carrier cartel, while (2) rail regulation has classic equalizing discrimination as its primary characteristic.

The removal of government oversight over rates and entry and the simultaneous outlawing of collective rate making should have diver-

[10] The simple correlation between cols. a and b is .71.

gent outcomes in the railroad and motor carrier industries. A deregulated trucking industry would—at least in truckload service—develop a rate structure more closely related to costs of service. A general reduction in trucking prices would be expected.

Deregulating rail rates would have far more complex results. Given the small number of rail carriers in any market, a strictly cost-based rate structure could not be predicted; nor, if rail marginal costs are below average costs, would this result be desirable. Instead, my analysis suggests that a more rational rail rate structure would result, one based both on costs and on the ability to divert traffic (the latter related, presumably, to trucking costs). Deregulating rail rates should not lead to a general increase or decrease in rail rates but to a reshuffling of prices.

This expectation is confirmed by the recent removal of ICC oversight over fruit and vegetable tariffs. The diversion of traffic to unregulated truckers over the last 2 decades effectively defeated the attempt at equalizing discrimination, and in 1979 rail rates on this traffic were deregulated. The result was a simultaneous increase of some rates and a sizable diversion of traffic to railroads in response to lower rates for other hauls (Association of American Railroads 1979; American Trucking Associations 1979). There is every reason to expect that deregulation of rail rates for other commodities would have a similar effect.

The disparate effects of regulation in the two modes are the direct result of the institution of pricing services individually rather than by categories. Individual pricing, which has allowed the ICC to pursue a policy of equalizing discrimination, has also required the use of rate bureaus and collective rate making. Had the transport industries used category rate making, regulation to enforce equalizing discrimination would have been hampered and collective rate making would have been unnecessary, and, therefore, it is doubtful that the fervent calls for and against deregulation would be heard today.

Appendix

The regressions in the text explain variations in fitted or standard rates for each commodity-origin-destination combination. These rates are adjusted for differences caused by variable loading weight among different commodities.

While most discussions of transport pricing proceed on the assumption that there is a single charge per unit of service, in fact this is not the case. There are separate charges per hundredweight for shipments of a commodity between two end points for (*a*) minimum charge, (*b*) less-than-truckload (LTL), (*c*) truckload (TL) or carload, and (*d*) very heavy shipments. Due to the different charging regimes there is an inverse relationship, "a weight taper," between charges per hundredweight and loading weight. To develop a

structure of single rates for commodities between end points, the first step was to estimate a weight taper for each commodity.

The weight tapers were estimated on the basic assumption that for every origin-destination pair, the shipping charges per hundredweight for a shipment of one size would be proportionate across different levels to the charges for a shipment of a standard vehicle load of the commodity. The standard procedure of spline estimation was used to estimate the shape of the taper in the absence of prior knowledge of its functional form. Knots were arbitrarily placed at 100 and 300 cwt for motor carrier estimation and at 500 and 1,000 cwt for railroads.

The functional form used to measure the rate taper with respect to loading weight for each commodity was:

$$\ln \text{cents}_i = b_1(\ln wt_i) + b_2(\ln wt_i)^2 + b_3(\ln wt_i)^3 + b_4F(\ln wt_i - \ln f)^3 + b_5G(\ln wt_i - \ln g)^3 + \sum_{j=1}^{n} c_jD_{ji} + e_i,$$

where

cents_i = cents per hundredweight for shipment;
wt_i = weight of shipment in hundredweight;
D_{ji} = haul dummies: $D_{ji} = 1$ if the shipment is on end-point pair j, $D_{ji} = 0$ if the shipment is not on end-point pair j;
F = 1 if $wt_i > 100$ for motor carrier or $wt_i > 500$ for railroads, 0 otherwise,
f = 100 for motor carriers, 500 for rail;
G = 1 if $wt_i > 300$ for motor carrier or $wt_i > 1{,}000$ for railroads, 0 otherwise;
g = 300 for motor carriers, 1,000 for rail; and
e_i = error term.

The values of the spline variables are listed in tables A1 and A2 for rail and motor carriers, respectively. The number of dummy variables is listed in each table as the number of hauls. For two commodities, there was insufficient variation in the independent variable to estimate the relationships for railroads.

In general, the tapers have at least one steeply downward-sloping segment which for motor carriers can be interpreted as the transition from LTL rates to TL rates. As should be expected from commodities that differ in loading characteristics, this steep downward-sloping section occurs at different weights for different commodities.

Some commodities have flat motor carrier tapers beyond the first downward-sloping segment, while others continue to fall; still others have a slightly upward-sloping segment at the very end Since spline functions are piecewise cubic, it is not clear how much faith should be placed in the aberrant behavior of the tapers of some commodities for very large and very small shipment sizes.

The downward-sloping segment of the weight taper for rail-shipped commodities is found in most cases at much larger shipment sizes than for motor carriers. For most rail commodities, the downward slope is not as steep as for trucks. As must be expected for commodities with different loading characteristics, there are exceptions to this rule. Without exception, however, the downward slopes for rail commodities are at a much higher range than for truck commodities.

TABLE A1

MOTOR CARRIER WEIGHT TAPER REGRESSIONS

Independent Variable	Commodity					
	2042190 Feed	2084120 Wine	2621490 Wrapping Paper	2631115 Pulpboard or Fiberboard	2651141 Corrugated Boxes	3011110 Tires
ln WT	−.61	−1.13	17.64	−8.92	−60.98	−14.61
$(\ln WT)^2$	.31	.57	−4.31	2.14	15.33	3.76
$(\ln WT)^3$	−.04	−.08	.34	−.18	−1.26	−.32
$(\ln WT - \ln f)^3$	.20	.55	−.60	.41	2.63	.77
$(\ln WT - \ln g)^3$	.38	−1.11	1.13	−.17	−4.33	−1.38
R^2	.996	.998	.989	.981	.931	.909
r^2	.781	.946	.605	.455	.56	.331
Observations	731	128	1,101	1,165	764	4,409
Hauls	644	108	923	958	632	3,308
F	57.34	49.43	52.79	33.64	4.66	108.42

SOURCE.—Rocky Mountain Motor Freight Bureau 1972

NOTE.—Dependent variable = ln (cents per hundredweight)

TABLE A2

RAIL WEIGHT TAPER REGRESSIONS

Independent Variable	Commodity			
	2042190 Feed	2621490 Wrapping Paper	2631115 Pulpboard or Fiberboard	3011110 Tires
b_1	172.48	−103.78	−128.61	−7 47
b_2	−30 02	17.61	21.13	1.10
b_3	1 72	−.99	−1.16	−.05
b_4	−3.81	1.85	1.62	−3.71
b_5	5.11	1.69	− 82	251 37
R^2	.963	998	.980	.979
r^2	.08	.90	378	296
Observations	1,187	562	3,159	999
Hauls	890	432	1,941	577
F	5 38	231	147 63	35 12

SOURCE —Interstate Commerce Commission 1972

NOTE —Dependent variable = ln (cents per hundredweight) Equations for wine and boxes could not be estimated due to insufficient variation in the independent variables

Two coefficients of multiple determination are reported for each equation. As usual, R^2 is interpreted as the percentage of the total variation in the dependent variable which can be explained by the right-hand-side variable. This is misleading, however, since the extensive use of dummy variables accounts for much of the apparent explanatory power of the regressions. The second coefficient of multiple determination, r^2, is defined as:

$$r^2 = \frac{\sum_{i=1}^{n_j}\sum_{j=1}^{m}(Y_{ji} - \bar{Y}_j)^2}{\sum_{i=1}^{n_j}\sum_{j=1}^{m}(Y_{ji} - \bar{Y}_j)^2 - \sum_{i=1}^{n_j}\sum_{j=1}^{m}(\hat{Y}_{ji} - \hat{\bar{Y}}_j)^2},$$

where Y_{ji} is the ith observation on the dependent variable in group j, $\bar{Y}_j$ is the mean of the dependent variable for observations in group j, and $\hat{Y}_{ji}$ is the fitted value of the ith observation on the dependent variable in group j. Numbers in the row marked r^2 are thus a more accurate indication of the explanatory power of the five spline variables. The reported F-statistic is computed using r^2.

The dummy variables in each equation were computed using the analysis of covariance. They were evaluated, however, not at 0 cwt but at 340 cwt for trucks and 680 cwt for trains. The dummy variables thus are interpreted as a standard rate for each haul and commodity: a least-squares forecast of the rate had all shipments been made at the standard size. The standard rate thus purges all influence of variable loading weights within a commodity. For the two rail commodities whose weight tapers could not be estimated, the standard rate was computed as the average charge for the haul, regardless of loading weight.

References

American Commercial Lines, Inc., et al. v. Louisville and Nashville Railroad Co. et al. "Ingot Molds Case." 392 U.S. 571 (1968).

American Trucking Associations. "Decontrolled Rail Movements of Fresh Fruits and Vegetables." *Res. Rev.*, no. 215 (October 15, 1979), pp. 2–3.

Association of American Railroads. "Railroads Reverse Decline in Fruit, Vegetable Traffic." *Information Letter*, no. 2280 (September 26, 1979), pp. 1–4.

Boyer, Kenneth D. "On the Structure of Regulated Freight Rates." Econometrics Workshop Paper no. 7801, Michigan State Univ., August 1978.

Council of Economic Advisers. *Economic Report of the President.* Washington: Government Printing Office, January 1979.

Friedlaender, Ann F. *The Dilemma of Freight Transport Regulation.* Washington: Brookings Inst., 1969.

Harris, Robert G. "Economics of Traffic Density in the Rail Freight Industry." *Bell J. Econ.* 8 (Autumn 1977): 556–64.

Hilton, George W. "The Basic Behavior of Regulatory Commissions." *A.E.R. Papers and Proc.* 62, no. 2 (May 1972): 47–54.

Levin, Richard C. "Allocation in Surface Freight Transportation: Does Rate Regulation Matter?" *Bell J. Econ.* 9 (Spring 1978): 18–45.

Locklin, D. Philip. *Economics of Transportation.* 7th ed. Homewood, Ill.: Irwin, 1972.

MacAvoy, Paul W., and Sloss, James. *Regulation of Transport Innovation: The ICC and Unit Coal Trains to the East Coast.* New York: Random House, 1967.

Meyer, J. R.; Peck, M.; Stenason, J.; and Zwick, C. *The Economics of Competition in the Transportation Industries.* Cambridge, Mass.: Harvard Univ. Press, 1959.

Rocky Mountain Motor Freight Bureau. "All-Bureau Continuing Traffic Study for 1972." Computer tape, Rocky Mountain Motor Freight Bur., Denver, 1972.

Spady, Richard H. *Econometric Estimation of Cost Functions for the Regulated Transportation Industries.* New York: Garland, 1979.

Suits, D. B.; Mason, A.; and Chan, L. "Spline Functions Fitted by Standard Regression Methods." *Rev. Econ. and Stats.* 60, no. 1 (February 1978): 132–39.

Williams, Winston. "Trucking Industry Mobilizes to Fight Rate Deregulations." *New York Times* (January 2, 1978), p. 32.

U.S. Interstate Commerce Commission. "One Percent Waybill Sample of Carload Terminations for 1972." Computer tape, ICC, Washington, 1972.

Transaction Costs, Order Placement Strategy, and Existence of the Bid-Ask Spread

Kalman J. Cohen and Steven F. Maier

Duke University

Robert A. Schwartz

New York University

David K. Whitcomb

Rutgers University

By considering investor order placement strategy, this paper demonstrates that transaction costs cause bid-ask spreads to be an equilibrium property of asset markets. With transaction costs, the probability of a limit order executing does not go to unity as the order is placed infinitesimally close to a counterpart market quote; thus, with certainty of execution at the counterpart market quote, a "gravitational pull" is generated that keeps counterpart quotes from being placed infinitesimally close to each other. An equilibrium spread is defined and its size linked to market thinness; implications are noted for the design of a trading system.

I. Introduction

This paper establishes that transaction costs in secondary asset markets cause individual investors to use order placement strategies that

We thank Amir Barnea, Avraham Beja, Richard Burton, Mark Eaker, Dan Galai, Kenneth Garbade, Thomas Ho, Wesley Magat, David Peterson, and Marshall Sarnat for their helpful comments. Earlier drafts of this paper have been presented at seminars in economics and finance at Duke University, the Hebrew University at Jerusalem, New York University, Tel-Aviv University, and the Joint National Meetings of TIMS/ORSA (New York City, May 1978)

[*Journal of Political Economy*, 1981, vol. 89, no. 2]
© 1981 by The University of Chicago 0022-3808/81/8902-0003$01.50

result in a nontrivial market bid-ask spread.[1] We define an *equilibrium* market spread and demonstrate that it will be greater for thinner securities.

The analysis fits into a growing body of literature which increasingly is being referred to as the microstructure of security markets. Stigler (1964), Demsetz (1968), West and Tinic (1971), Tinic (1972), and Tinic and West (1972, 1974), along with Farrar, Smidt, Stoll, and others involved in the institutional investor study (see U.S. Securities and Exchange Commission 1971), were among the first to focus rigorous analytical attention on the operations of security markets. More recently, microstructure theory has been explicitly considered by, among others, Garman, who coined the term (1976), Beja and Hakansson (1977), and Cohen, Maier, Schwartz, and Whitcomb (hereinafter CMSW) (1978). The analytical issues addressed involve the interplay among market participants, trading mechanisms, and the dynamic behavior of security prices. In the present paper, we study the formulation of optimal investor trading strategies and how these interact with one aspect of the dynamic behavior of security prices, the bid-ask spread. Spreads are of concern to investors because they are a variable cost of trading by market order and because they cause an inflation of transaction-to-transaction returns variance.

The pioneering analysis of bid-ask spreads was provided by Demsetz (1968). Further studies include: Tinic (1972), Tinic and West (1972, 1974), Benston and Hagerman (1974), Garman (1976), Hamilton (1976, 1978), Branch and Freed (1977), Stoll (1978*a*, 1978*b*), Ho and Stoll (1979, 1980), Newton and Quandt (1979), Schleef and Mildenstein (1979), Smidt (1979), and Amihud and Mendelson (1980). In CMSW (1979) we discuss this literature and contrast it with the formulation presented here. Except for Ho and Stoll (1980), there has been no explicit consideration of the transition from individual spreads to the market spread. It will be clear from our analysis that this transition is not a simple aggregation process and that the market spread is the product of a dynamic interaction involving many market participants. Also, previous theoretical models have generally assumed that investors can be dichotomized into two groups—immediacy demanders and immediacy suppliers. Our model of investor order placement strategy suggests that such a dichotomy will not be observed in the marketplace.

For the market, the spread is the difference between the lowest ask and the highest bid of all participants. In markets composed of many

[1] We have elsewhere (Cohen et al. 1980) considered how the impact of transaction costs on stock price movements introduces serial correlation in returns data and causes estimates of the market model beta coefficient to be biased.

traders with heterogeneous beliefs and trading propensities, one might expect to have orders at virtually every permissible price in the neighborhood of equilibrium and hence to find no significant market spread.[2] However, we show that even when expectations and trading propensities are heterogeneous, the spread is a property of asset markets that have temporarily cleared.[3] The analysis yields an existence proof of a noninfinitesimal spread with continuous pricing; a fortiori this proves for discrete prices the existence of a positive spread for nontrivial reasons.[4]

The essence of our argument is as follows. At any point in time, any investor might alternatively seek to trade via a limit order (be an immediacy supplier), trade with certainty via a market order (be an immediacy demander), or not seek to trade at all. Limit orders create the book, and market orders clear out limit orders. Because execution via market order is certain, while execution via limit order is not, it never pays for any investor to place a limit order (e.g., a bid) at a price too close to that of a counterpart limit order (e.g., an ask). Intuitively stated, as a trader contemplates placing a bid closer and closer to an ask already established on the market, he is increasingly attracted by this counterpart offer; at some point, the "gravitational pull" exerted by the established ask will dominate. The trader will "jump" his price and execute with certainty via a market order.

Section II establishes the scenario for our analysis. Section III focuses on the probability of a limit order executing and shows that, with transaction costs, this probability does not go to unity as the price at which a limit order is placed becomes infinitesimally close to a counterpart market quote. This demonstration underscores an investor's need for an order placement strategy and provides the foundation for our gravitational pull model. Section IV models the investor's order placement decision and develops conditions under which he will transmit limit or, alternatively, market orders to the market, or do nothing. The analysis is developed in a dynamic programming framework, although we are interested in the descriptive modeling of

[2] When an asset market has cleared, there is neither excess demand nor excess supply in the sense that at that moment no market participant is willing to buy the asset at a price equal to or greater than the ask, and no one is willing to sell at a price equal to or less than the bid. Hence we must have a market spread that is at least equal to the minimum allowable price change for the asset in question (on the major U.S. stock exchanges, this is 1/8 of a dollar for most common stocks). However, it is not obvious why spreads greater than minimum allowable price changes are commonly observed.

[3] While the analysis presented here is applicable to any asset market, our formal model treats the secondary market for financial securities. Security markets have two convenient properties: All units of an asset are identical, and such markets are impersonal (which means that bargaining, as distinct from trading, strategies need not be considered).

[4] See n. 2 above.

an investor's decision process rather than in actually generating a normative solution. Section V demonstrates that implementation of the strategic order placement decision (which implies the gravitational pull effect) causes a noninfinitesimal bid-ask spread to exist. This section also defines an equilibrium bid-ask spread, discusses conditions under which it will exist, and shows that it is positively related to market thinness. Section VI considers implications for the design of a security market trading system and summarizes our analysis.

II. The Scenario

Consider an investor who maximizes the expected utility of terminal wealth and, for simplicity, let him allocate funds between only two assets: a risk asset and cash (which we take to be the numeraire asset). In the absence of transaction costs, the market would be monitored continuously and appropriate transactions would be made with each change in the market price and the investor's demand propensities. Then, if price were continuous, there would be no spread and the market price would be determined by a straightforward aggregation of individual demand propensities.

However, a variety of transaction costs impact on the investor's trading decisions. The fixed (with respect to number of shares traded) costs of assessing information, monitoring the market, and conveying orders to the market imply that the investor will make trading decisions only periodically. Further, when decisions are made, he will not convey his full set of demand propensities to the market. For one thing, trades that involve sufficiently small portfolio adjustments would not justify the transaction costs incurred.[5] Also, attempts to transmit several limit orders simultaneously would be likely to overload our current system. Furthermore, a continuous auction which does not generate a Walrasian solution cannot readily handle multiple buy-sell orders that, ex ante, are intended to be alternatives.[6]

In light of transaction costs (and also taking account of the timing and magnitude of exogenous cash flows), the investor will establish a discrete set of decision points. In the analysis presented in Section IV, we take the frequency of these points as predetermined. Upon reaching a decision point, an investor can do nothing, or hit an

[5] It can be readily demonstrated that variable as well as fixed transaction costs make sufficiently small portfolio adjustments prohibitively expensive.

[6] That is, if any array of buy and sell orders from one investor is executed sequentially, the desired quantities at each price should be dependent on the exact sequence of purchases and sales followed. However, the investor does not know ex ante which specific sequence will occur. This problem could, of course, be handled (at a cost) by conditioning orders on the sequence of prior transactions (i.e., if the limit order to buy 200 at $50 executes, then sell 100 at $56).

existing limit order with his own market order, or place his own limit order at a "better" price and run the risk of its not executing. A concurrent strategy issue also exists when the investor finds it prohibitively expensive to convey his entire set of demand propensities to the market; for all intents and purposes, he must select the single best alternative to transmit.

III. The Probability of a Limit Order Executing

In this section, we establish the conditions under which the probability of execution does not approach a limit of unity as the price at which the limit order is placed is taken to be infinitesimally close to the counterpart market quote. Under these conditions we obtain a probability jump (at the counterpart market quote) that underlies the gravitational pull effect developed in Section V. By showing that the probability jump can be attributed to the existence of transaction costs, we establish the basic linkage between spreads and transaction costs.

Consider the case where an investor contemplates submitting a limit bid at the price P_t^{LB} at time t, and let price be a continuous variable. A similar argument can be constructed for the case of a limit ask. Assume that if the limit order is unfilled by the next decision point at time $t + 1$ the order will be canceled. Let L be the length of time between decision points t and $t + 1$.

Let P_t^{MA} be the market ask price at time t. Consistent with the random walk version of the efficient markets hypothesis, we make the Markov assumption that each subsequent market ask depends only on the last previous market ask. If we also assume that each change in the natural log of the market ask, Z_t, is a random variable that is independently and identically distributed over time with mean zero and variance var (Z_t), then we can model the market ask price generation process as a compound Poisson stochastic process:[7]

[7] For expositional simplicity we assume that the stochastic processes considered in this section have no drift (that is, their expected value is zero) It should be noted that even though $\ln P_t^{MA}$ is assumed to have no drift, the price series itself, P_t^{MA}, may have drift For example, when $\ln (P_{t+1}^{MA}/P_t^{MA})$ is normal, with mean zero and standard deviation σ, then P_{t+1}^{MA}/P_t^{MA} is log normal with mean exp $(\frac{1}{2}\sigma^2)$.

There will be a realization of the random variable Z_t at each point of time when any one of the following events occurs to affect the specific limit order which sets the market ask. (*a*) it is withdrawn; (*b*) it executes against a crossing buy order; or (*c*) it remains on the book but is no longer the market ask since a lower limit sell has been submitted Note that events of types *a* and *b* necessarily result in a change in the market ask when price is continuous and utility functions are heterogeneous (which implies a zero probability of having two or more orders at a specific price). Also note that only events of type *b* are associated with transactions; hence the number of Z_t which materialize during any interval will usually exceed the number of transactions in that interval.

$$\ln P_t^{MA}(\Delta) = \ln P_t^{MA} + \sum_{i=1}^{N(\Delta)} Z_i, \tag{1}$$

where Δ is the time from the last decision point. When Δ equals L, we have $P_t^{MA}(\Delta) = P_{t+1}^{MA}$; when Δ equals $2L$, we have $P_t^{MA}(\Delta) = P_{t+2}^{MA}$, etc. The number of changes in the market ask that take place in the time interval Δ is $N(\Delta)$. We assume $N(\Delta)$ follows a Poisson process with arrival rate ν.

The next step is to determine the probability of execution (during a time interval of length L) of a limit bid which is submitted at a price P_t^{LB} greater than the current market bid (P_t^{MB}) but less than the current market ask (P_t^{MA}). Clearly, as the limit bid price approaches (from below) the market ask, the probability of execution increases. One might suppose that this probability approaches unity as the limit bid approaches the market ask; however, this need not be the case. In the Appendix we restate formally and prove:

PROPOSITION 1: If P_t^{MA} is generated by the compound Poisson process of equation (1), then no matter how close the limit bid approaches (from below) the market ask, the probability of the limit order executing is less than unity in any time interval of finite length.

Since a market order will always execute with probability one, proposition 1 gives a probability jump at the market ask.

Transaction costs are crucial to the existence of the probability jump. In the absence of transaction costs, one might expect that, following the work of Merton (1973), the logarithm of the market ask would best be described by a Wiener process with zero drift:

$$d \ln P_t^{MA}(\Delta) = \sigma dZ(\Delta), \tag{2}$$

where σ is the instantaneous variance of the process and $dZ(\Delta)$ is a standardized Wiener process. In this case, the price $P_t^{MA}(\Delta)$ would experience an infinite number of adjustments in the interval $0 \leq \Delta \leq L$. In the Appendix we restate formally and prove:

PROPOSITION 2: If $P_t^{MA}(\Delta)$ is generated by the Wiener process of equation (2), then as the limit bid approaches (from below) the market ask, the probability of the limit order executing approaches unity for all time intervals of the finite length.

Proposition 2 implies that for the Wiener process there will be no probability jump at the market ask.

We next consider whether there is a relationship between the compound Poisson process of equation (1) and the Wiener process of equation (2). In the Appendix we restate formally and prove:

PROPOSITION 3: If the random variable Z_i in the compound Poisson process of equation (1) has two equally likely possible values, $+\alpha$ and $-\alpha$, and if the arrival rate ν of this compound Poisson process increases without bound while α simultaneously decreases in such a way that $\nu\alpha^2$ is constant, then the compound Poisson process approaches the Wiener process described by equation (2) with the variance of the Wiener process, σ^2, equaling $\nu\alpha^2$.

Proposition 3 states that the compound Poisson process can be expected to approach the Wiener process under appropriate assumptions.[8] Thus the probability jump of proposition 1 would disappear as ν increased and var (Z_i) decreased. In the Appendix we restate formally and prove:

PROPOSITION 4: As the arrival rate ν in the compound Poisson process of equation (1) increases, the probability of a limit order executing increases for all $P_t^{LB} < P_t^{MA}$.

Hence the probability of a limit order executing increases at all $P_t^{LB} < P_t^{MA}$ as the activity proxy ν increases. Stated conversely, the "thinner" a security, that is, the less active are investors in submitting orders to trade in the security, the lower will be the probability of execution at each $P_t^{LB} < P_t^{MA}$, and therefore the greater will be the probability jump at P_t^{MA}.

The four propositions stated in this section (and proved in the Appendix) have important implications for the analysis of the impact of transaction costs on bid-ask spreads. Without transaction costs, the market price could be expected to behave as a Wiener process; that is, there would be an infinite number of infinitesimally small price adjustments. The investor who was considering placing a limit order could reduce the probability of his order not executing to as close to zero as desired simply by placing his order close enough to the counterpart market quotation.

On the other hand, with transaction costs and a finite number of investors, the market price would generally not behave as a Wiener process. Investors would find continuous adjustments in their portfolios too expensive, and market prices would behave as a stochastic jump process (proposition 3 shows that this process could generally be expected to approach a Wiener process as the order arrival rate is increased). One such jump process is the compound Poisson, and it also is consistent with a (martingale) efficient market.

[8] The particular distribution chosen for Z_i in proposition 3 is not critical. Other appropriately chosen discrete or continuous distributions can also be shown in the limit to go to the Wiener process.

Proposition 1 demonstrates that for such a stochastic jump process a probability jump exists. Proposition 4 has shown that the probability jump will be greater for thinner securities. In Section V we will show that the larger probability jump for thinner issues leads to equilibrium market spreads which are larger for thinner issues.

IV. A Model of Investor Order Placement Strategy

We now consider the question of when an investor will choose to trade via a market order or limit order, or not seek to trade at all. The problem is structured as follows. Because of the costs of monitoring the market, let the investor consider rebalancing his portfolio only at preselected points in time, $t = 1, 2, \ldots, T - 1$, where T is the investment horizon.[9] In order to simplify the analysis, we now consider the placement of only purchase orders for the risk asset (omitting the symmetric case of sell orders without loss of generality).

At any of the $T - 1$ decision points, the investor is faced with three possible courses of action:

a) Submit a market order to buy shares at the current market ask price of P_t^{MA}.

b) Submit a limit order to buy shares at a limit bid price of $P_t^{LB} < P_t^{MA}$.

c) Do nothing.

In modeling the investor's choice among these three alternatives, we find it convenient to make the following assumptions that simplify the analysis without materially affecting the nature of the conclusions. We assume that all orders are for a fixed number of shares ΔN and that when any market or limit order is executed, it is satisfied fully at the stated price; this avoids the tedium of writing (average) transaction price and transaction costs as functions of the number of shares exchanged and of defining probabilities of partial execution We assume that unfilled limit orders are canceled prior to the next decision point; this avoids both the need to include additional state variables (the price and quantity of any limit order outstanding) and the need to analyze additional courses of action (submit a market order and leave the old limit order outstanding, or submit a market order and remove the limit order, etc.). Finally, we assume no lags in the transmission of information and orders; this avoids the complexity of dealing with changes in the current market quotes during the time the investor formulates and implements his decision.

[9] Note that the rebalancing points need not be the same for all investors, so that trades can occur at any time when the market is open even if specific traders are not continually in the market.

If option a is chosen, the market order will be executed, and the investor's holdings at time $t + 1$ become $N_{t+1} = N_t + \Delta N$, for which the investor pays a total cost of $\Delta N \cdot P_t^{MA} + C^M$ where C^M is the total cost of transmitting and executing the market order.

If option b is chosen, then one of two events can take place: ($b1$) The limit order is executed. The investor's share holdings then become $N_{t+1} = N_t + \Delta N$, for which the investor pays a total cost of $\Delta N \cdot P_t^{LB} + C^{L1} + C^{L2}$ where C^{L1} is the cost of transmitting a limit order to the market and C^{L2} is the cost of executing a limit order. ($b2$) The limit order does not execute and is canceled prior to the next decision point. The investor's share holdings then remain unchanged ($N_{t+1} = N_t$) and his cash is decreased by the cost of transmitting the limit order, C^{L1}. Option b will be chosen over a if the gain associated with the possibility of trading at a more favorable price outweighs the loss associated with the probability of not trading at all.

The investor must consider four subjective probability distributions in order to make an optimal decision. These are: (1) the joint probability distribution of market bid and ask prices at time $t + 1$, conditional upon the quotes at time t; (2) the probability of a limit bid order submitted at time t executing before time $t + 1$; (3) and (4) the joint probability distribution of market bid and ask prices at time $t + 1$, conditional upon the quotes at time t and further conditional on whether a limit bid submitted at time t either did or else did not execute prior to time $t + 1$. In the Appendix we discuss in more detail these four subjective probability distributions (only three of which are independent).

A dynamic programming model of the investor's choice among options a, b, and c is formulated in the Appendix. The investor is assumed to maximize his expected utility of terminal wealth, max (U_1, U_2, U_3), where U_1 is the expected terminal utility of choosing option a, trading via a market order; U_2 is the expected terminal utility of choosing option b, seeking to trade via a limit order; and U_3 is the expected terminal utility of choosing option c, doing nothing. It is convenient to focus on the utility gain, ΔU_1 or ΔU_2, which results from choosing option a or b rather than option c: $\Delta U_1 = U_1 - U_3$; $\Delta U_2(P_t^{LB}) = U_2(P_t^{LB}) - U_3$ (note that U_2 and ΔU_2 are functions of the limit bid price). Clearly $\Delta U_2(P_t^{LB} = P_t^{MA}) = \Delta U_1$, since, at this price, the market order and limit order strategies are effectively the same (the probability is unity of the limit order executing at a price of P_t^{MA}).

Let us now consider the conditions under which $U_1 > \max(U_2, U_3)$, in which case the investor will submit a market order, or $U_2 > \max(U_1, U_3)$, in which case the investor will submit a limit order. Suppose that at the current market quotes the do-nothing strategy is dominated (i.e., $\max[U_1, U_2] > U_3$). Given our utility gains ΔU_1 and

$\Delta U_2(P_t^{LB})$ and the probability function for limit order execution developed in Section III above, we know the utility gain from placing a market order and can readily obtain an expected utility gain function for the limit order strategy. These are depicted in figure 1; the shape of the function is explained as follows:

1. While the utility of a consummated trade decreases monotonically with P_t^{LB}, the probability of execution increases with P_t^{LB}, with two probability jumps (at the market bid and the market ask). The jump at P_t^{MB} simply reflects the institutional reality that orders placed at prices less than or equal to P_t^{MB} would not have priority over the current market bid, whereas an order placed at a price above P_t^{MB} would be at the top of the limit order book. The jump at P_t^{MA} follows from proposition 1.

2. Since the probability of execution is constant to the right of P_t^{MA}, the expected utility gain has a peak at P_t^{MA}, corresponding to the utility gain of transacting by a market order.

3. The probability of execution increases between P_t^{MB} and P_t^{MA}, with the greatest relative increase just to the right of P_t^{MB}. We expect this large probability increase in this neighborhood because the strategy considerations of other investors might lead them to place limit orders just to the right of P_t^{MB} in order to capture the largest price advantage. Hence the second peak in the expected utility gain function will occur at some point P'', between P_t^{MB} and P_t^{MA}.

4. The probability of execution decreases rapidly in a neighborhood just to the left of P_t^{MB} because of existing limit orders on the book. However, since there would be a clustering of limit orders near

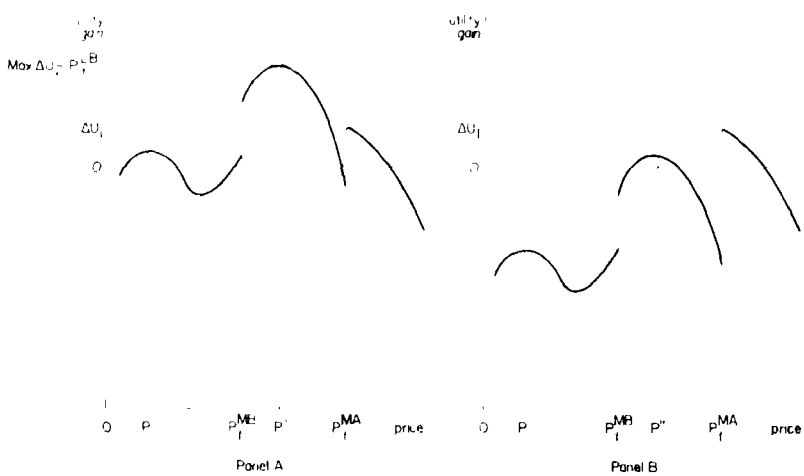


FIG. 1.—Illustrative utility gain functions

P_t^{MB}, we would expect the probability of execution to decrease more slowly as the limit price, P_t^{LB}, moves further from P_t^{MB}. Hence, we might find a third peak in the expected utility gain function to occur at some point P' to the left of P_t^{MB}.

Which of the three peaks in $\Delta U_2(P_t^{LB})$ will dominate depends on the particular form of the probability function for limit order execution as well as the expected utility function. In the particular diagram presented in figure 1*A*, $\Delta U_2(P_t^{LB})$ has positive values (hence a do-nothing strategy is dominated), has values greater than ΔU_1 (hence the market order strategy is dominated), and gives a global maximum at P''. Assuming only one limit order per investor, the optimal strategy is to place a limit bid at P'', and hence P'' will become the new market bid. The setting of such a bid quotation can be considered a Stigler (1961)-type search activity.[10]

Clearly the limit order strategy need not always be superior; in selecting this strategy, the investor also accepts the chance that the limit order may fail to execute, in which case the investor would be worse off than had he done nothing since he would have also lost the cost of placing the limit order, C^{L1}. If the probability of failing to execute is high enough and $\Delta U_1 > 0$, the investor will prefer the market order strategy as illustrated in figure 1*B*.[11]

V. The Market Spread

It is clear from the preceding analysis (Sec. IV) that, with a continuous auction market, each investor's order placement decision is made with reference to prices already established on the market—the bid and ask quotations which define the market spread. In turn, each investor's order may affect the market spread to which subsequent traders react. Hence, the market spread is the product of a dynamic interactive process. In this section we show that a nonzero market spread must exist, define the equilibrium market spread, and relate a security's equilibrium spread to its thinness.

The limit order book comprises the limit orders transmitted to the market by a subset of the many traders in a security, and the spread is essentially a gap in the limit order book. Therefore, having estab-

[10] By the placement of such a limit order, a seller or buyer announces his propensity to trade with the hope of getting the attention of a counterpart market participant who would also be willing to trade

[11] There are two other versions of fig. 1. One would be analogous to panel *A*, but with the global maximum at P' rather than at P'': in this case, the investor would submit a limit bid at P' (below the market bid). The second version would be where the zero point on the utility gain axis is higher than both max $\Delta U_2(P_t^{LB})$ and ΔU_1, in this case, the investor would do nothing.

lished that individual investors will sometimes seek to trade via limit orders, we must show why a noninfinitesimal gap in the array of such orders exists between the market quotes. We do so by considering the impact that the jump in the probability of a limit order executing (Sec. III) has on the investor's optimal order placement strategy (Sec. IV).

We have established the conditions under which a limit order will be placed; we can now consider whether or not limit orders will necessarily be placed so as always to preserve a nonzero bid-ask spread. By continuity of the utility function, and because of the discontinuous drop in the probability of execution that must occur at the market ask as we move to lower prices, we must have a discontinuous drop in $\Delta U_2(P_t^{LB})$ as we move past the market ask to lower prices. Then, because of continuity of the probability function until we get to the market bid (at which point another discontinuous drop occurs), it necessarily follows that $\Delta U_2(P_t^{LB})$ cannot exceed ΔU_1 within a nonzero neighborhood to the left of P_t^{MA}. Hence, no limit bid will be placed infinitesimally close to the market ask, and therefore we must have a nonzero bid-ask spread.

Intuitively viewed, consider an investor who is thinking about whether or not to place a limit bid at some price below the market ask. As the potential limit price rises toward the market ask, it becomes relatively more attractive to transact at the market ask with certainty. At some point, the investor will prefer buying at a discretely higher, unitary probability of execution. In other words, certainty of execution at the market ask creates a kind of gravitational pull which causes investors to jump their price once their potential limit bid gets close enough to the market ask. This clears out limit bids when they get too close to the market ask, thereby leaving a nonzero spread between the market bid and ask.

We next consider the issue of an equilibrium market spread.

DEFINITION: In a dynamic trading process we define the equilibrium market spread as the bid-ask spread at which, for the next instant of time, the probability of the spread increasing is equal to the probability of the spread decreasing.[12]

This definition does not imply that if away from equilibrium the

[12] Note that the probability is not conditioned on the current status of individual investors, recent prices, or other market conditions.

Note that we use the term "equilibrium market spread" to refer to one particular point in the probability distribution of the market spread, rather than to refer to a statistic for the central tendency of the entire distribution. We do not adopt a central tendency definition (such as expected value, mode, or median) of the equilibrium spread because it does not provide a ready link to the market forces that generate the spread and to thinness. For empirical research, the average market spread is likely to be a reasonable proxy for our definition of the equilibrium market spread.

spread will necessarily move toward it on the next order, but only that it is more likely to do so than not.

We assume that a unique equilibrium exists. To examine the reasonableness of this assumption, first consider an arbitrarily small spread. Because the probability of a limit order being placed has a discrete jump at P_t^{MA}, we have already argued that there will be no limit orders placed in some nonzero neighborhood below P_t^{MA} by *all* investors seeking to buy the security. (The generalization from a single investor to all investors takes place simply by choosing the smallest neighborhood found among all investors.) Therefore, for a sufficiently small bid-ask spread, the unconditional probability of the spread increasing must exceed the unconditional probability of the spread decreasing as analyzed from the standpoint of the buyer. A similar argument holds for the seller.

Now consider what occurs when the spread widens. The larger the spread is, the greater the potential utility gain from an optimally priced limit order, while the utility gain from a market order will remain unchanged or fall. The gain in utility occurs because the shares can be bought (sold) at a lower (higher) price via limit order than by market order.[13] With the rise in utility $U_2(P_t^{LB})$, investors will begin to shift their preferences from market orders to limit orders. As more and more investors shift to limit orders, the probability of the spread increasing falls, while the probability of it decreasing rises.

One of two situations must occur. Either the spread would tend to reach a point at which the two probabilities would be equal, in which case it would be the equilibrium market spread, or the limit order book would be empty on one or both sides.[14] Since when the order book is empty market orders cannot transact, we would be at a point where the probability of a limit order is greater than that of a market order. Although such a point might be achieved without the equilibrium market spread ever being attained, we believe it would be atypical.[15]

[13] We assume that the spread does not impart any information to the investor on the direction of future price movements.

[14] No other possibility exists. So long as there is a greater probability that the spread will increase than that it will decrease, the spread will on expectation grow—limited only by the collapse of the limit order book on one side or the other.

[15] With regard to the assumption of uniqueness, investors might have utility functions which could lead to more than one equilibrium market spread. Consider the case where an investor views a very wide spread as signaling the advent of new information which may cause a price adjustment of unknown direction. The investor sensing greater risk might now choose a do-nothing strategy over placing either a limit or market order. This could support a very wide spread as a new market equilibrium spread. Of course, investors would eventually conclude that the new information was not forthcoming, and trading would resume at its former frequency with the old equilibrium spread reestablished.

Our definition of the equilibrium market spread provides a direct link between the market spread and thinness. Recall that proposition 4 showed that the probability of a limit order executing decreases, for all values of P_t^{LB}, as a security's order arrival rate decreases. Thus, ceteris paribus, thin issues have a lower probability of limit orders executing than do thick issues. This lower probability in turn decreases the proportion of investors choosing limit orders over market orders at any given size of the spread; this implies that, for a thinner issue, a wider spread would be required for the forces that increase and decrease the spread to be in equilibrium. Hence we have:

PROPOSITION 5: Thinner securities will, ceteris paribus, have larger equilibrium market spreads.

VI. Conclusion

We have presented a proof of the existence of market spreads in markets with many traders. The formulation treats continuous prices and allows for heterogeneous trading propensities. The proof is not dependent on the demand of traders for immediacy or on the cost to market makers of providing immediacy.

The literature on bid-ask spreads does not appear to have recognized that aggregation from individual to market spreads is a considerably more complex process than the standard aggregation from individual to market demand and supply functions. Neither has it been established that ordinary investors may sometimes seek to trade via limit orders (and at other times via market orders), hence that these investors will sometimes supply (and at other times demand) immediacy, and that in choosing between market and limit orders investors implement an order placement strategy. In addressing these issues, we have sought to establish the links between transaction costs, individual investor order placement strategy, market thinness, and market spreads.

We have first established that, with transaction costs, the probability of a limit order executing does not rise to unity as the price at which the order is placed gets infinitesimally close to a counterpart market quote. We have next shown that the resulting investor trading strategies generate what we have referred to as a gravitational pull effect. Essentially, in the neighborhood of the current market bid and ask quotations, what might otherwise have been limit orders are instead submitted as market orders (at slightly less desirable prices) so as to achieve certainty of execution. These market orders trigger trades which clear limit orders off the book, widening the market spread. The gravitational pull effect explains why market spreads may be

substantial even in markets composed of many traders. Finally, we have defined an equilibrium market spread (where the forces that tend to widen and to narrow the spread are in balance) and have shown it to be positively related to a security's thinness (measured inversely by the order arrival rate).

Our formulation has several implications for the design of a market system. A primary objective of system design should be to expand the extent and frequency with which investors interact with the market by minimizing various transaction costs. Decreasing variable transaction costs will decrease individual spreads and generate a greater order flow; decreasing the costs of monitoring and communicating with the market will also increase the frequency with which investors rebalance their portfolios; and consolidating the currently fragmented system (by, e.g., instituting a consolidated limit order book) will reduce search costs and further shrink spreads by increasing the effective thickness of the market. These costs are all a function of market structure and hence should be amenable to reduction by appropriate system design. However, a major cost of interacting with the market is the cost of decision making, and this might not be subject to significant reduction by exchange organization. For this reason, it is possible that, especially for thinner issues, spreads will remain sizable in a restructured national market system.

Appendix

In this Appendix, we restate formally and prove propositions 1–4 of Section III above and develop the more technical aspects of the arguments presented in Section IV.

Section III

Equation (1) of Section III presents a compound Poisson process as the stochastic process which generates a sequence of market ask prices over time. This can be used to determine the probability that a limit bid order will execute during a particular time period.

Suppose that the price of the limit order to be submitted is greater than the current market bid but less than the current market ask, that is, that $P_t^{MB} < P_t^{LB} < P_t^{MA}$. To determine the probability that the potential limit bid will execute in a time interval of length L, we must find the probability that $P_t^{MA}(\Delta)$ will decrease to a price equal to or less than P_t^{LB}.[16] Therefore, let $Y(L, P_t^{MA}, P_t^{LB})$ be the probability that the minimum value that $P_t^{MA}(\Delta)$ achieves in the interval $0 \leq \Delta \leq L$ is equal to or less than P_t^{LB}. As P_t^{LB} approaches P_t^{MA}, we would expect the probability of the potential limit bid executing to increase, since the amount that $P_t^{MA}(\Delta)$ would have to decline would be reduced. However, in the

[16] For P_t^{MA} discretely greater than P_t^{LB}, the possibility of another investor submitting a limit order with a price greater than P_t^{LB} will only tend to decrease the probability of execution of the original limit order. Thus our proof of existence of the probability jump at P_t^{MA} is conservative

limit as P_t^{LB} approaches P_t^{MA} from below, without further analysis it is unclear how far the probability will rise. Therefore, let

$$\phi(L, P_t^{MA}) = 1 - \lim_{P_t^{LB} \to P_t^{MA}} Y(L, P_t^{MA}, P_t^{LB}),$$

where the limit is understood to be from below. We now prove:

PROPOSITION 1: If $P_t^{MA}(\Delta)$ is generated by the compound Poisson process of equation (1), then $\phi(L, P_t^{MA}) > 0$ for all intervals of length $L < \infty$.[17]

PROOF: Since Z_i is stochastic with mean zero, $P(Z_i \geq 0) > 0$. Furthermore, the value of the Poisson random variable $N(L)$ will be finite in any interval of length L less than infinity. Therefore, the probability of $N(L)$ consecutive Z_i observations that are greater than or equal to zero is given by $[P(Z_i \geq 0)]^{N(L)}$, which, since $N(L)$ is finite, must be strictly greater than zero. Notice that if all the Z_i observations are greater than or equal to zero, the value of $P_t^{MA}(\Delta)$ must be greater than or equal to P_t^{MA} throughout the interval $0 \leq \Delta \leq L$; thus the limit order P_t^{LB} would not have executed. This is sufficient to demonstrate the probability jump. Clearly, there are other sample paths that $P_t^{MA}(\Delta)$ could have followed which also would have failed to execute the limit order.

PROPOSITION 2: If $P_t^{MA}(\Delta)$ is generated by the Wiener process of equation (2), then $\phi(L, P_t^{MA}) = 0$ for all intervals of length $L < \infty$.

PROOF: By the reflection principle for a continuous Wiener process (see Karlin 1968, pp. 276–77) and since $\ln P_t^{MA}(\Delta)$ is driftless,

$$\begin{aligned} Y(L, P_t^{MA}, P_t^{LB}) &= \Pr\{\min_{0 \leq \Delta \leq L} P_t^{MA}(\Delta) \leq P_t^{LB}\} = \Pr\{\min_{0 \leq \Delta \leq L} \ln P_t^{MA}(\Delta) \leq \ln P_t^{LB}\} \\ &= 2 \Pr\{\ln P_{t+1}^{MA} < \ln P_t^{LB}\} = 2 \Pr\{P_{t+1}^{MA} < P_t^{LB}\} \\ &= \frac{2}{\sigma\sqrt{2\pi L}} \int_{-\infty}^{P_t^{LB}} \exp\left\{-\frac{1}{2L}\left(\frac{x - P_t^{MA}}{\sigma}\right)^2\right\} dx, \end{aligned}$$

where the latter probability distribution follows from the definition of the Wiener process. Now

$$\lim_{P_t^{LB} \to P_t^{MA}} Y(L, P_t^{MA}, P_t^{LB}) = \frac{2}{\sigma\sqrt{2\pi L}} \int_{-\infty}^{P_t^{MA}} \exp\left\{-\frac{1}{2L}\left(\frac{x - P_t^{MA}}{\sigma}\right)^2\right\} dx.$$

Substituting the variable $y = (x - P_t^{MA})/(\sigma\sqrt{L})$, the preceding limit equals $(2/\sqrt{2\pi}) \int_{-\infty}^{0} \exp\{-\frac{1}{2} y^2\} dy = 2(\frac{1}{2}) = 1$.

PROPOSITION 3: If the random variable Z_i is expressed as a Bernoulli random variable, with $Pr(Z_i = \alpha) = Pr(Z_i = -\alpha) = 1/2$, and if the arrival rate ν of the Poisson process $N(\Delta)$ goes to infinity, while simultaneously reducing the size of α in such a way that $\nu\alpha^2$ remains constant, the compound Poisson process approaches the Wiener process described by equation (2) with $\sigma^2 = \nu\alpha^2$.

PROOF: This follows from the theorem that the characteristic function for the compound Poisson approaches that of the Wiener process (see Parzen 1962, p. 99).

PROPOSITION 4: $[\partial Y(L, P_t^{MA}, P_t^{LB})]/\partial\nu > 0$ for all $P_t^{LB} < P_t^{MA}$ and $L < \infty$.

PROOF: By the Markov property of the compound Poisson process, increasing (decreasing) the order arrival rate by some factor λ is identical to

[17] This proposition would also hold for other stochastic jump processes where the number of price changes is finite in any finite interval.

increasing (decreasing) the length of time L between decision points by λ. When λ is increased, those sample paths of the process which initially satisfied

$$\min_{0 \leq \Delta \leq L} P_t^{MA}(\Delta) \leq P_t^{LB}$$

will still continue to do so. On the other hand, some sample paths where

$$\min_{0 \leq \Delta \leq L} P_t^{MA}(\Delta) > P_t^{LB}$$

will now satisfy the inequality

$$\min_{0 \leq \Delta \leq \lambda L} P_t^{MA}(\Delta) \leq P_t^{LB}.$$

Therefore, $Y(L, P_t^{MA}, P_t^{LB})$ would increase for $\lambda > 1$ for all values of $P_t^{LB} < P_t^{MA}$. Similarly, $Y(L, P_t^{MA}, P_t^{LB})$ would decrease if the order arrival rate λ is decreased.

Section IV

We now present the more technical aspects of the model of an investor's order placement strategy discussed in Section IV The subjectively determined probability distributions are specified as follows. First, the probability distribution of future market bid and ask prices is given by $h(P_{t+1}^{MA}, P_{t+1}^{MB} \mid P_t^{MA}, P_t^{MB})$, where h is a joint density function for market asks and bids $(P_{t+1}^{MA}, P_{t+1}^{MB})$ at time $t + 1$, given the prices at t. Consistent with random walk models of security price behavior, we condition future prices only on current prices. Consistent with Section III, we let the investor's subjective probability of a limit order executing before $t + 1$ be given by $p(P_t^{LB}, P_t^{MA}, P_t^{MB})$, where P_t^{LB} is the limit bid price. (Again, this is consistent with random walk theory, since we need know only current market quotes to predict whether or not the limit order will be executed.) Last, we assume the investor determines conditional probability distributions of future prices in the event of either a successful or unsuccessful limit order. That is, for a limit bid at price P_t^{LB} submitted at time t, let $k(P_{t+1}^{MA}, P_{t+1}^{MB} \mid P_t^{MA}, P_t^{MB}, P_t^{LB})$ be the joint density function of market bid and ask prices at time $t + 1$ if the limit bid executes prior to $t + 1$, and let $l(P_{t+1}^{MA}, P_{t+1}^{MB} \mid P_t^{MA}, P_t^{MB}, P_t^{LB})$ be the joint density function at time $t + 1$ if the limit bid fails to execute prior to $t + 1$. Note that only three of the four subjective probability distributions h, p, k, and l can be independently determined, since by Bayes's theorem we must have

$$\begin{aligned} h(x, y \mid P_t^{MA}, P_t^{MB}) = {} & p(P_t^{LB}, P_t^{MA}, P_t^{MB}) k(x, y \mid P_t^{MA}, P_t^{MB}, P_t^{LB}) \\ & + [1 - p(P_t^{LB}, P_t^{MA}, P_t^{MB})] l(x, y \mid P_t^{MA}, P_t^{MB}, P_t^{LB}) \end{aligned}$$

for all possible values of P_t^{MA}, P_t^{MB}, and P_t^{LB}.

We can now give a formal statement of the model. Given observed market quotes of P_t^{MA} and P_t^{MB} and holdings by the investor of N_t shares of the security and S_t dollars in cash, at time t the investor's expected utility of terminal wealth can be written as $\psi_t = f_t(P_t^{MA}, P_t^{MB}, N_t, S_t)$.[18]

[18] Note that the investor's underlying utility function measures the utility of the wealth he will possess at some horizon T, where $T > t$. The investor will choose those decisions at times $t, t + 1, \ldots, T - 1$ which maximize the expected utility of his terminal wealth. As of time t (the investor's current decision point), viewing the expectations operator as ranging over relevant random variables pertaining to times between t and T, we define ψ_t to be the expected utility of the investor's terminal wealth Clearly this depends upon the assets the investor has at time t (N_t and S_t) and the current market prices at which the investor can buy or sell shares (P_t^{MA} and P_t^{MB}).

We now derive an expression for ψ_t in terms of the various probability assessments and costs. Write $\psi_t = \max(U_1, U_2, U_3)$, where U_1 = expected terminal utility if a market order is placed at time t, U_2 = expected terminal utility if a limit order is placed at time t, and U_3 = expected terminal utility of doing nothing at time t. More precisely, letting x and y be values for the ask and bid prices at time $t + 1$, respectively, we have

$$U_1 = \int_0^\infty \int_0^\infty f_{t+1}(x, y, N_t + \Delta N, S_t - C^M - N \cdot P_t^{MA}) \cdot h(x, y \mid P_t^{MA}, P_t^{MB}) dx dy.$$

As stated in Section IV, we assume that the number of shares to be purchased, ΔN, is fixed. Define the function

$$\begin{aligned} U_2(P_t^{LB}) = {} & p(P_t^{LB}, P_t^{MA}, P_t^{MB}) \int_0^\infty \int_0^\infty f_{t+1}(x, y, N_t + \Delta N, \\ & S_t - C^{L1} - C^{L2} - \Delta N \cdot P_t^{LB}) \cdot k(x, y \mid P_t^{MA}, P_t^{MB}, P_t^{LB}) dx dy \\ & + [1 - p(P_t^{LB}, P_t^{MA}, P_t^{MB})] \int_0^\infty \int_0^\infty f_{t+1}(x, y, N_t, S_t - C^{L1}) \\ & \cdot l(x, y \mid P_t^{MA}, P_t^{MB}, P_t^{LB}) dx dy \end{aligned}$$

for all possible prices of the limit order. Then we have

$$U_2 = \max_{P_t^{LB}} U_2(P_t^{LB}).$$

Finally, for the do-nothing option we have

$$U_3 = \int_0^\infty \int_0^\infty f_{t+1}(x, y, N_t, S_t) \cdot h(x, y \mid P_t^{MA}, P_t^{MB}) dx dy.$$

The dynamic programming recursion permits us to obtain a solution for f_t in the order $T - 1, T - 2, \ldots, 2, 1$. This is possible since we assume the utility value for f_T is known for all values of the parameter P_T^{MA}, P_T^{MB}, N_T, and S_T. We have defined this recursion in terms of the four state variables P_t^{MA}, P_t^{MB}, N_t, and S_t. We also have two decision variables, P_t^{LB} and the decision as to which of the three courses of action to take.

References

Amihud, Yakov, and Mendelson, Haim. "Dealership Market: Market-making with Inventory." *J. Financial Econ.* 8 (March 1980): 31–53.

Beja, Avraham, and Hakansson, Nils H. "Dynamic Market Processes and the Rewards to Up-to-Date Information." *J. Finance* 32 (May 1977): 291–304.

Benston, George J., and Hagerman, Robert L. "Determinants of Bid-Asked Spreads in the Over-the-Counter Market." *J. Financial Econ.* 1 (December 1974): 353–64.

Branch, Ben, and Freed, Walter. "Bid-Asked Spreads on the Amex and the Big Board." *J. Finance* 32 (March 1977): 159–63.

Cohen, Kalman J.; Hawawini, Gabriel A.; Maier, Steven F.; Schwartz, Robert A.; and Whitcomb, David K. "Implications of Microstructure Theory for Empirical Research on Stock Price Behavior." *J. Finance* 35 (May 1980): 249–57.

Cohen, Kalman J.; Maier, Steven F.; Schwartz, Robert A.; and Whitcomb,

David K. "The Returns Generation Process, Returns Variance, and the Effect of Thinness in Securities Markets." *J. Finance* 33 (March 1978): 149–67.

———. "Market Makers and the Market Spread: A Review of Recent Literature." *J. Financial and Quantitative Analysis* 14 (November 1979): 813–35.

Demsetz, Harold. "The Cost of Transacting." *Q.J.E.* 82 (February 1968): 33–53.

Garman, Mark B. "Market Microstructure." *J. Financial Econ.* 3 (June 1976): 257–75.

Hamilton, James L. "Competition, Scale Economies, and Transaction Cost in the Stock Market." *J. Financial and Quantitative Analysis* 11 (December 1976): 779–802.

———. "Marketplace Organization and Marketability: NASDAQ, the Stock Exchange, and the National Market System." *J. Finance* 33 (May 1978): 487–503.

Ho, Thomas, and Stoll, Hans R. "Optimal Dealer Pricing under Transactions and Return Uncertainty." Working Paper, New York Univ., Graduate School Bus. Admin., 1979.

———. "On Dealer Markets under Competition." *J. Finance* 35 (May 1980): 259–67.

Karlin, Samuel. *A First Course in Stochastic Processes.* 2d ed., enl. New York: Academic Press, 1968.

Merton, Robert C. "An Intertemporal Capital Asset Pricing Model." *Econometrica* 41 (September 1973): 867–87.

Newton, William, and Quandt, Richard E. "An Empirical Study of Spreads." Working Paper, Princeton Univ., Dept. Econ., 1979.

Parzen, Emanuel. *Stochastic Processes.* San Francisco: Holden-Day, 1962.

Schleef, Harald J., and Mildenstein, Eckhard. "A Dynamic Model of the Security Dealer's Bid and Ask Prices." Paper presented at meetings of the Western Economic Association, Las Vegas, 1979.

Smidt, Seymour. "Continuous vs. Intermittent Trading on Auction Markets." Working Paper, Cornell Univ., Graduate School Bus. and Public Admin., 1979.

Stigler, George J. "The Economics of Information." *J.P.E.* 69, no. 3 (June 1961): 213–25.

———. "Public Regulation of the Securities Markets." *J. Bus.* 37 (April 1964): 117–42.

Stoll, Hans R. "The Supply of Dealer Services in Securities Markets." *J. Finance* 33 (September 1978): 1133–51. (*a*)

———. "The Pricing of Security Dealer Services: An Empirical Study of NASDAQ Stocks." *J. Finance* 33 (September 1978): 1153–72. (*b*)

Tinic, Seha M. "The Economics of Liquidity Services." *Q.J.E.* 86 (February 1972): 79–93.

Tinic, Seha M., and West, Richard R. "Competition and the Pricing of Dealer Service in the Over-the-Counter Stock Market." *J. Financial and Quantitative Analysis* 7 (June 1972): 1707–27.

———. "Marketability of Common Stocks in Canada and the U.S.A.: A Comparison of Agent versus Dealer Dominated Markets." *J. Finance* 29 (June 1974): 729–46.

U.S. Securities and Exchange Commission. *Institutional Investor Study Report.* 92d Cong., 1st sess. House Document no. 92–64, 1971.

West, Richard R., and Tinic, Seha M. *The Economics of the Stock Market.* New York: Praeger, 1971.

Land Value Capitalization in Local Public Finance

David A. Starrett

Stanford University

We explore the conditions under which the welfare benefits of local public goods projects will be capitalized into land values. We find two types of sufficient conditions, one involving similar communities and the other involving differentiated communities. The form of capitalization differs between these cases, and we explore the nature of these differences. We also examine intermediate cases and identify models in which there will be no capitalization.

I. Introduction

The idea that the value of local public goods projects will be capitalized into land rents has a long tradition in the public finance literature, dating back at least as far as Tiebout (1956). Empirical evidence has been mixed, but some positive results have been reported by Oates (1969), among others. This simple idea is potentially very powerful. It provides a simple measure of project benefits which reveals (if perhaps after the fact) individual preferences for public goods. And if rent changes can be predicted in advance, it may provide an objective function for community decision making.

Furthermore, the presence of capitalization forces has important equity implications. To the extent that capitalization occurs, benefits tend to accrue to landlords at the expense of renters. We will show that capitalization reflects a real welfare increase only if these two groups are considered equally deserving. If, on the other hand, renters are considered more deserving, then the presence of capitalization clearly will worsen the distribution of benefits within society.

[*Journal of Political Economy*, 1981, vol. 89, no. 2]
© 1981 by The University of Chicago 0022-3808/81/8902-0009$01.50

The "capitalization hypothesis" is interesting from a purely theoretical point of view as well. It is clear that increases in land rent values are not themselves a net social benefit; owners of the land will benefit, but renters will suffer. Indeed, in the standard surplus theory of welfare measurement, changes in land prices (or prices in any other competitive market) never appear as net benefits (in aggregate, the gains to winners just offset losses to losers). Thus, if capitalization occurs, it must happen because the true benefits induce a corresponding change in rents.

We will argue that there are two separate and independent forces which can induce capitalization, one involving competitive forces between communities and one involving such forces within a particular community. Furthermore, these two forces have quite different implications for (1) what benefits are capitalized and (2) where those benefits are capitalized. Thus, we will derive several different capitalization theorems and explore the conditions under which each holds. In the process, we will find some conditions under which capitalization does not occur at all or occurs only partially.

Let us refer to capitalization which derives from forces between communities as "external capitalization" and capitalization which derives from forces within communities as "internal capitalization." The intuitive argument for external capitalization is quite simple. Suppose that a project is built in one community designed to make people better off there. Now, if people in all communities have similar tastes and are free to move among the communities, outsiders will necessarily be attracted to the project-building community. And they will continue to move until the welfare incentive disappears. The only factor which can stop this movement is a differential location cost, that is, an increase in land rents in the project-building community. This type of argument is the one used by Polinsky and Shavell (1976) to justify a form of capitalization. However, the Polinsky/Shavell model is not very general, since the benefits accrue to absentee landlords. We will show below that increases in land values generally will overstate true net social benefit, although they may be a good approximation to community net benefit.

Naturally, when we drop the assumption of free mobility between communities or the assumption of similar tastes between communities, the intuition for external capitalization is weaker. Indeed, several authors have argued that we should not expect land value capitalization in specialized (or Tiebout) communities which are immune to migration forces due to differences in tastes (see Hamilton 1975 and Polinsky and Rubinfeld 1978).

However, there is a quite different argument for internal capitalization which is based on specific properties of local public goods.

Presumably public goods are local in character when proximity matters. For example, one has to make trips in order to appreciate such public goods as parks, museums, civic centers, sports complexes, and even roads and highways. Now, if there is a positive correlation between the amounts of public goods and the differential desirability of various locations, some degree of capitalization will occur as rents adjust. We will show later exactly what conditions are necessary for full capitalization. This type of argument has been used in the literature to justify the use of rent gradients for measuring the cost of pollution near a source (such as an airport or factory).[1] It is also in the spirit of arguments for capitalization given by a number of authors, among them Strotz (1968), Lind (1973), and Pines and Weiss (1976). These authors essentially start with the assumption that a project will differentially improve land quality and explore whether or not those improvements will translate into rent increases. The results are instructive, but the underlying approach seems to beg the question somewhat. The primary question is, Will a local public project change the differential quality of land? For the case of irrigation projects discussed by Lind, the answer is clearly yes, but in other cases it is less obvious. Indeed, we will show that, for some types of public goods, no capitalization occurs at all.

We will want to pay attention to the system of taxation in discussing both types of capitalization. One might guess that the form and nature of capitalization will depend on whether or not the public goods are financed out of a property tax. We will show that this is sometimes, but not always, the case.

We adopt a somewhat generalized version of the Polinsky/Shavell community model in the sequel. This model has a number of special features but is chosen because it facilitates the exposition and makes for easy comparisons with the preceding literature. However, many of the results reported here remain valid in a much more general economic model. The interested reader is referred to Starrett (1978) for some general results on capitalization.

II. A Prototype Model

Throughout this section we deal with a model in which there is a single local public good (q) being produced independently by a number of communities. Further, we will suppose that projects initiated in a community will not affect private goods prices other than land rents.

[1] For an example of this argument in the literature, see Freeman 1971.

This may seem a serious restriction, but it is not, really, since changes in other prices will have the usual canceling effects on the two sides of the market. With nonland private goods prices fixed, we can aggregate all these goods into a single commodity and let this commodity be the numeraire in what follows.

Agents must engage in effort (take trips) to enjoy the public good. Naturally, such trips are costly; here we will assume that the expense can be represented by a simple cost function in terms of the numeraire: $f(g,s)$, where g is the number of trips taken and s stands for location within the community. Location will be specified by dividing the community's land into discrete zones and assuming that location can adequately be specified by designating the appropriate zone.

Each consumer gets to choose a zone and an amount of land within the zone (ℓ). The market for land will be assumed competitive, and r_s will stand for the rental rate in zone s.

Consumers earn income in terms of the numeraire. We will want to separate this income into three terms: profit shares (Π), rental income on land (R), and other income (mainly labor) (Y). Thus, an individual in income class i has income $I^i = \Pi^i + Y^i + R^i$. Such an individual will be required to pay taxes T^i. The T^i may be a function of other parameters, depending on the nature of taxation.

A consumer of preference type a has a preference function of the form $U^a[q,g,\ell,I^i - T^i - f(g,s) - r_s\ell]$. The arguments of this function are the level of public goods provision, the number of trips taken, the amount of land consumed, and the amount of other private goods consumed. It seems appropriate that both g and q should affect utility. For example, if the public good is a museum, agents will care about the size of the museum as well as the number of visits made.

Some of the results reported below will depend on the properties of U^a and f. Specific assumptions will be introduced as we go along, but the following general discussion may be helpful here. It seems reasonable that the marginal cost of trips ($\partial f/\partial g$) will increase as the distance from the public good increases. We expect to find that use decreases with distance if and only if this assumption holds. Also, it is reasonable to suppose that the marginal benefit from public goods ($\partial U/\partial q$) increases with use (g). This marginal benefit will decrease with distance whenever both assumptions hold.

Each consumer in a particular community gets to choose g, ℓ, and s. (At a later stage such a consumer may be able to choose a community of residence as well.) We will assume throughout that there is free mobility within the community, so that s is a free choice. It is analytically convenient to think of each resident as making a conditional

choice of g and ℓ for each s. This process generates a first-stage optimization problem:

$$\max_{g,\ell} U^a[q,g,\ell,I^t - T^t - f(g,s) - r_s\ell]. \tag{1}$$

A solution to problem (1) defines demand functions $g_s^{at} \equiv g^a(q,I^t - T^t,r_s,s)$ and $\ell_s^{at} \equiv \ell^a(q,I^t - T^t,r_s,s)$ and an indirect utility function $V_s^{at} \equiv V^a(q,r_s,I^t - T^t,s)$. At a later "stage," the consumer makes a discrete location choice, seeking to $\max_s V_s^{at}$. We will use the notation V^{at} to denote the final indirect utility function.

The remainder of the community economy consists of firms, about which we need to say little except that they behave competitively and are not affected by the public goods, and the government, which produces the public good from private goods, paying for those goods with revenues from the taxes levied on households. Letting $\Gamma(q)$ stand for the public goods cost function, the government must satisfy a balanced budget condition $T = \Gamma(q)$, where T stands for total tax revenue.

We assume that the private sector is always in market equilibrium. For nonland commodities, this simply means that markets clear at fixed (unchanging) prices. For land, it means that rental rates must adjust until the market for land clears in each zone.

We will formulate questions about capitalization as follows. We measure welfare using a standard Bergson-Samuelson formulation. We ignore distributional considerations by assuming that the distribution of income is initially optimal (such an assumption is clearly necessary for pure forms of capitalization). Next, we consider an initial situation and a proposed new project. Some form of capitalization occurs when there is a correlation between the resulting welfare change and the associated rental change. We will see that this correlation can take many forms.

All precise statements about internal capitalization require some boundary condition on community rents. This boundary condition will take different forms, depending on the structure of the town and the organization of taxation in the town. We will discuss two different types of communities in this regard. In the first type of community the boundaries will be considered potentially variable, with "outside" land always available to the community at an exogenously given opportunity cost; the boundary rent is thus exogenously given in such a community. The real prototype for such a community would be a town or metropolitan area located in the middle of farmland. We will refer to such communities as isolated.

In the second type of community, the boundaries are predetermined (presumably by legal arrangement). The effect of a project on

boundary rent will be determined by changes (if any) in the aggregate demand for land in the community. The real prototype here would be a subcommunity within a larger metropolitan area. We will refer to such communities as adjacent.

The local public goods model being proposed here seems most applicable to the case of isolated communities, since it is difficult to ignore the importance of direct spillovers among adjacent communities. However, it still seems useful to treat the adjacent case.

To demonstrate that there are separate forces generating external and internal capitalization, we will need to show that it is possible to have each without the other. We will begin with a discussion in which we assume that no external forces are present; that is, we assume that community policy will not lead to any net migration to or from the community. Within this context, we explore the conditions for internal capitalization. We derive alternative sets of assumptions under which there is no capitalization or full capitalization.

Then we will return to the issue of external forces. We will give these forces their best chance to work by assuming free mobility between communities. To isolate the effect of these forces, we will adopt a model of the community which is consistent with no internal capitalization. Within this context, we again exhibit conditions under which there is either no capitalization or full capitalization.

A. *Internal Capitalization*

Assuming away migration effects, there are essentially two ways in which project benefits can get translated into increased land values: through a change in the extensive margin or the intensive margin. If a project increases the aggregate demand for land, it may lead to an increase in the general level of rents (by operating on the extensive margin for land). Intuitively, this force could go either way. When taxes are raised (to pay for a new project), there is likely to be a substitution away from goods (including land). On the other hand, when public goods provision is increased, there is likely to be a complementary increase in the demand for land. Since the net effect on rents is clearly ambiguous, there is no reason to expect systematic capitalization from impact on the extensive margin. In fact, we will argue that the extensive margin is likely to be of no importance, especially in isolated communities.

However, the project may affect the differential value of locations and thereby influence the rent structure through the internal margin. We will show that this force may lead to systematic capitalization in relatively large communities, within which there is a wide range of choice regarding use of the public goods. Restricting this range of

choice can lead to any intermediate case from full capitalization to no capitalization.

B. Sufficient Conditions for No Internal Capitalization

The essential restriction which is needed to rule out internal capitalization is that there be no effective differential choice in the use of public goods within the community. A number of sets of assumptions will do for this purpose.

The easiest case to explain is one in which there is no differential choice within the community at all. By this, we mean that $g = \bar{g}$ and $\ell = \bar{\ell}$ for all agents in all zones; these conditions would hold if the plot size is institutionally given and the public good has the characteristics of national defense so that residents get the benefits regardless of their actions.

A resident of preference type a and income class i now chooses s (the only remaining choice variable) to $\max_s U^a[q,\bar{g},\bar{\ell},I^i - T^i - f(\bar{g},s) - r_s\bar{\ell}]$. Clearly, this problem reduces for every agent to one of minimizing costs: $\min_s [f(\bar{g},s) - r_s\bar{\ell}]$. Since the costs to be minimized are the same for all agents, the only possible equilibrium rent structure must be one which makes those costs independent of s. (At such a cost structure, everyone is obviously indifferent as to where they live.) Now, this independence condition must be true both before and after the project is initiated. It follows that the derivative of costs with respect to q must be independent of s. But with g and ℓ fixed, this means that $(d/dq)r_s$ must be independent of s; that is, if rents change at all they must change in a uniform way. However, rents at the boundary cannot change: There is no change in the aggregate demand for community land, so the boundary rent will not change in either isolated or adjacent communities. Hence there is no capitalization. Thus we can identify one case which formalizes the assertions of Hamilton (1975).

We can introduce further choice into the community and obtain the same results as long as we are willing to restrict the degree of diversity and the degree of complementarity of tastes. For example, suppose that tastes are separable and similar in that preferences of all agents can be represented in the form $U = \hat{U}^1(q,g) + \hat{U}^2(\ell) + I^i - T^i - f(g,s) - r_s\ell$, where i indexes incomes class, as before. Now, if it should turn out that the optimal choice of g is independent of s (so that there is no effective differential choice within the town), we can again argue that there will be no capitalization. Clearly agents still make the same choices regardless of income class. Thus, rents must still adjust so that everyone is indifferent as to where they live (V^i_s is independent of s). This condition must be true both before and after the project, so

$(d/dq)V_s^i$ must be independent of s. Differentiating and using the envelope theorem, this condition becomes that

$$\frac{d}{dq}\hat{U}(q,g_s) + \ell_s \frac{dr_s}{dq} \tag{2}$$

is independent of s. But since we assumed that g_s was independent of s, the first term of (2) is independent of s, and we are back to the statement that the rent structure must change uniformly if at all. In fact, it cannot change at all, because there is no change in the extensive demand for land (given separable preferences).[2]

C. *Sufficient Conditions for "Full" Internal Capitalization*

Internal capitalization can take a number of different forms, depending on specific assumptions made concerning the boundary conditions, the type of taxes imposed, and the nature of property ownership. We will delineate these cases as we go along. As will be seen, all of the propositions reported here rely exclusively on the intensive land margin to "enforce" capitalization. There may be some very special functional forms which will generate systematic capitalization from the extensive margin, but these would seem too special to be of much interest.

There are two principal assumptions which taken together guarantee that project benefits will be translated to land rents through the intensive margin. The first is that all the benefits of the project must be "intramarginal" in that boundary residents are marginally unaffected. We would expect this condition to hold if marginal residents choose $g = 0$ (so that there is a complete revealed range of choice within the community). But the condition they may hold more generally; for example, in the case of a museum or park, "boundary" residents may make so few trips that, at their current level of activity, they could make no better use of a larger facility. We formalize this condition as follows. Let σ stand for some boundary region. Then "marginal indifference" will mean that

$$\frac{\partial}{\partial q} U[q,g_\sigma^i,\ell_\sigma^i,I^i - T^i - f(g,\sigma) - r_\sigma \ell_\sigma^i] = 0. \tag{3}$$

The second necessary assumption is that residents do not tend to sort themselves out within the town according to their relative prefer-

[2] This case is closest in spirit to the example of no capitalization given in Polinsky and Shavell (1976). It also corresponds to Hamilton's (1975) model of Tiebout communities in which there is no internal spatial structure. As we show below, Hamilton's conclusion that there would be no capitalization in Tiebout communities depends critically on the assumption of no internal spatial structure.

ence for the public good. Since systematic differences will generally lead to sorting, this means that the community must be reasonably homogeneous in attitudes toward public goods. We will see later that, when sorting occurs, it will tend to mitigate capitalization.

The required homogeneity will result whenever agents are additively separable in their preferences on any component of consumption concerning which they systematically differ. Here we will treat the case where individuals differ in incomes and have preferences which are additively separable in the associated numeraire consumption good.

Consequently, we now assume that a resident (i) of the community has preferences which can be represented in the form $U^i = \hat{U}(q,g,\ell) + I^i - T^i - f(g,s) - r_s\ell$. Now, it is obvious that residents will not sort themselves out by income class. Conditional on living at s, all residents would make the same choices of g and ℓ ($g_s^i = g_s$, $\ell_s^i = \ell_s$, all s), so at any specified rent structure, they would all agree on the best location; hence, as before, the rent structure must adjust until all locations are equally desirable for all residents.

We now evaluate the first-order welfare effect of a new project:

$$dW = \sum_i \omega_i dV^i. \tag{4}$$

Since we are ignoring distributional factors, we must assume that whatever income differences prevail are "optimal," implying that the welfare weights (ω_i) are equal (otherwise, pure transfers could be made in such a way as to improve net welfare). We normalize units by setting the common weight equal to one. In evaluating (4) we treat everyone as if they were living at the boundary region σ and take into account all potential effects which q could have on the parameters faced by consumers.

Performing the differentiation, applying the envelope theorem where appropriate, and aggregating where possible, we have

$$\frac{dW}{dq} = N\frac{\partial}{\partial q}\hat{U}(q,g_\sigma,\ell_\sigma) - N\ell_\sigma\frac{dr_\sigma}{dq} + \frac{dY}{dq} + \frac{d\Pi}{dq} + \frac{dR}{dq} - \frac{dT}{dq},$$

where N stands for the total population of the community and the income variables without superscripts stand for aggregates over all residents. Now, marginal indifference (3) means that the first term is zero, and dY/dq must be zero as well, since Y cannot change under the assumption that nonland private goods prices are fixed. Furthermore, government budget balance implies that $dT/dq = d\Gamma/dq$. Making these substitutions, we have, finally,

$$\frac{dW}{dq} = \frac{dR}{dq} - \frac{d\Gamma}{dq} + \frac{d\Pi}{dq} - N\ell_\sigma\frac{dr_\sigma}{dq}. \tag{5}$$

The formula (5) can be thought of as a generic capitalization result from which a variety of specific relationships can be derived.

1. Local Ownership

Let us first look at the case where local firms (as well as local land) are all owned by local residents. In that case Π should be thought of as total profits generated in the region. Assuming that local firms behave competitively, we know that $d\Pi$ can be written in the form $d\Pi = X \cdot \delta p$, where X is the vector of net outputs and p the private goods price vector. But the only prices which change are land prices, so we have $d\Pi/dq = -(dR_f/dq)$, where R_f is the value of land used by firms. Substituting this relationship into (5), we have

$$\frac{dW}{dq} = \frac{dR_r}{dq} - \frac{d\Gamma}{dq} - N\ell_\sigma \frac{dr_\sigma}{dq}, \tag{6}$$

where R_r is the value of residential land.

Alternatively, we could take the view that firms exhibit constant returns to scale and are constantly in long-run equilibrium. In this case, the above analysis is invalid since any infinitesimal change in prices will lead to a discrete jump in the firm's decisions; in particular, if rents go up, ceteris paribus, all firms go out of business. The correct analysis would then set $d\Pi = 0$ in (5) and conclude (as in most of the previous examples in the literature) that capitalization involves all land value. Upon reflection, it seems that the long run is too long to be relevant here. Firms are not free to leave when rents rise, and the increasing rent payments by firms should be counted against profits. Hence we will stand by the formulation (6).

The final form capitalization takes depends on the boundary conditions. In the isolated case, it seems reasonable to suppose that $dr_\sigma/dq = 0$ even though the boundary of the town may shift. Conceptually, we can think of there being many boundary regions which are shared by residents and farmers. As long as some of these regions are still shared after the project is initiated, boundary rent cannot change. Thus for the case of isolated communities, we assert $dW/dq = (dR_r/dq) - (d\Gamma/dq)$. Verbally, the gross benefits of the project are capitalized into residential land values. Net benefits (dW/dq) are gross benefits minus costs $(d\Gamma/dq)$.

A more general statement can be made which will apply to adjacent communities as well. If the aggregate demand for land increases so that $dr_\sigma/dq > 0$, then the value of residential land overcapitalizes gross benefits ($[dR_r/dq] - N\ell_\sigma[dr_\sigma/dq]$ exactly capitalizes gross benefits) while if $dr_\sigma/dq < 0$, then the value of residential land undercapitalizes gross benefits.

2. National Ownership

The majority of firms are not locally owned. Indeed, most firms are national in the scope of their operations, and local ownership has no meaning for such firms. Suppose we take the position that all ownership is in national firms. What happens to the analysis? Actually, not very much. Indeed, the formula (6) is still correct for society as a whole, but now some of the costs of the project are paid by "foreigners" who own shares in the firms which are operating locally. Indeed, if we take the position that each community is small relative to the country, then it is a good approximation to assume that local branches of national firms are owned entirely by outsiders. In that case, it follows that gross benefits to residents of the community are capitalized into total land value, while gross benefits to society as a whole are capitalized into residential land values. The difference represents an externality. We will discuss other examples of such externalities later.

3. Property Taxes

Up until now we have dealt with only the case in which taxes were treated as lump-sum taxes by consumers. Since the major locally imposed tax is a property tax, it is important to see how the analysis needs to be modified in that case. If an ad valorem tax at rate t is imposed, then an agent using an amount of land ℓ in zone s will pay taxes $tr_s\ell$, where r_s must now be thought of as the rent net-of-tax payments.

Now, the problem for a typical individual can be stated as $\max_{g,\ell,s}$ $[\hat{U}(q,g,\ell) + I^i - f(g,s) - r_s(1+t)\ell]$. Clearly, it is still true that all agents will make the same choices, and, indeed, the method of analysis is exactly as before. Performing the calculations (and taking the point of view of society in order to avoid any ambiguity concerning the ownership of firms), we derive

$$\frac{dW}{dq} = \frac{dR_r}{dq} - Nr_\sigma\ell_\sigma\frac{dt}{dq} - N\ell_\sigma(1+t)\frac{dr_\sigma}{dq}, \tag{7}$$

where R_r now stands for total residential rent net of taxes.

The exact form of capitalization again depends on boundary conditions. And now there is some ambiguity concerning these even in the case of isolated communities. The issue revolves around whether farmers in the boundary regions do or do not pay the taxes. If they do pay the taxes, then the free boundary condition is as before. However, if they do not pay the taxes (but would have to if they acquired

property "within" the community), then the appropriate condition is $r_\sigma(1 + t) = \bar{r}$, where $\bar{r}$ is the opportunity cost of land to a farmer. We will refer to this arrangement as *agricultural zoning* and to the other case as *blind zoning*.

The case of agricultural zoning is easiest to analyze. Differentiating the boundary condition with respect to q, we have $(1 + t)(dr_\sigma/dq) + r_\sigma(dt/dq) = 0$. Therefore, referring back to (7),

$$\frac{dW}{dq} = \frac{dR_r}{dq}. \tag{8}$$

Here net project benefits are capitalized into net-of-tax residential property values, while gross project benefits are capitalized into before-tax property values.

For the case of blind zoning, the term involving dr_σ/dq vanishes in (7), and we are left to evaluate $Nr_\sigma\ell_\sigma(dt/dq)$. Suppose that only residential land is taxed, so that the government budget balance condition is $tR_r = \Gamma$. Differentiating this equation with respect to q and substituting for dt/dq in (7) yields

$$\frac{dW}{dq} = \frac{dR_r}{dq} - \alpha\left(\frac{d\Gamma}{dq} - t\,\frac{dR_r}{dq}\right) = (1 + \alpha t)\,\frac{dR_r}{dq} - \alpha\,\frac{d\Gamma}{dq}, \tag{9}$$

where $\alpha = N\bar{r}\ell_\sigma/R_r$. The net benefit is now a weighted difference between increases in after-tax residential property values and increases in cost. Note that the smaller is α, the closer we approximate the results for the previous case. The α stands for the ratio of the value of land residents would use if they all lived at the boundary to the actual value of residential land.

More generally, however, (9) tells us that increases in residential property value overcapitalize net benefits by an amount related positively to the gap between extra costs and the extra revenue that would be generated without changing tax rates. Clearly, this result is intermediate between pure gross capitalization and pure net capitalization.

D. *Preference Differences within the Community*

Once systematic preference differences are introduced into the communities, the exact capitalization results tend to break down. Here, we demonstrate this and indicate the type of modification which is required. For this purpose, let us consider the simplest extension possible, one with two different types of residents in the town (indexed a and b). Let there be N_a (N_b) a-type (b-type) residents. It is convenient to assume that all a-type residents have the same preferences and income (although different incomes could be incorporated as before).

Also, we find it necessary to suppose that the plot size within any given zone must be uniform, and we may as well assume that this plot size is predetermined.

Now we let S^a stand for the set of zones occupied by a-type residents and S^b for the corresponding set of zones for b-type residents. Obviously if σ is included in both sets, the analysis is exactly as in previous sections; indeed, all previous theorems go through with preference differences as long as those differences do not lead to systematic sorting. Therefore, without loss of generality, we can assume that the zoning arrangements are as in the Venn diagram shown in figure 1.

Since the zoning is discrete, there is little loss of generality in assuming that at least one zone is occupied by both types. We let τ stand for such a zone.

In what follows, it is convenient to use the shorthand notation $F_s^x(q) = \hat{U}^x(q,g_s^x,\ell_s) - f(g_s^x,s)$, all s, $x = a,b$. Thus, V_s^a can be written in the form $V_s^a \equiv F_s^a(q) - r_s\ell_s + I^a$. Next we evaluate welfare as before, except that now we cannot evaluate everyone's welfare at σ; instead, we evaluate a-type welfare at σ and b-type welfare at τ:

$$\begin{aligned} W &= N_a V_\sigma^a + N_b V_\tau^b, \\ &= N_a F_\sigma^a + N_b F_\tau^b + I - N_a r_\sigma \ell_\sigma - N_b r_\tau \ell_\tau. \end{aligned} \tag{10}$$

Finally, we make use of the fact that a-type agents are indifferent between locations τ and σ. This indifference implies that

$$r_\tau \ell_\tau = r_\sigma \ell_\sigma + F_\tau^a(q) - F_\sigma^a(q). \tag{11}$$

Substituting (11) into (10) yields $W = NF_\sigma^a(q) + N_b[F_\tau^b(q) - F_\tau^a(q)] + I - Nr_\sigma\ell_\sigma$, and differentiating with respect to q, we obtain:

$$\begin{aligned} \frac{dW}{dq} &= N\frac{d\hat{U}^a}{dq}(q,g_\sigma^a,\ell_\sigma) - N\ell_\sigma\frac{dr_\sigma}{dq} + \frac{d\Pi}{dq} + \frac{dR}{dq} - \frac{dT}{dq} \\ &\quad + N_b\left[\frac{d\hat{U}^b}{dq}(q,g_\tau^b,\ell_\tau) - \frac{d\hat{U}^a}{dq}(q,g_\tau^a,\ell_\tau)\right]. \end{aligned} \tag{12}$$

Examining (12), we see that the terms are just as before except for the intramarginal benefit term:

$$N_b\left[\frac{d\hat{U}^a}{dq}(q,g_\tau^b,\ell_\tau) - \frac{d\hat{U}^a}{dq}(q,g_\tau^a,\ell_\tau)\right]. \tag{13}$$

We will argue that there is a strong presumption that this term is positive. To begin with, the condition $g_\tau^b > g_\tau^a$ is a stability condition for the town structure we have specified. It guarantees that there is a differential benefit to an a-type agent relative to a b-type agent as one moves toward the boundary of the town (b-type agents gain relatively more from proximity to the center since they make more trips).

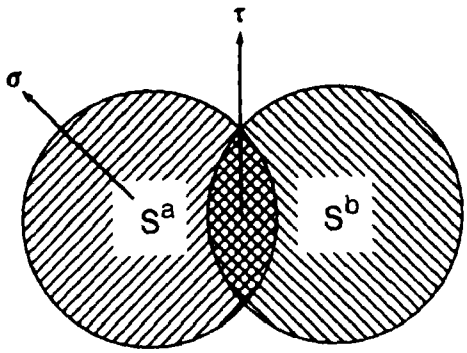


FIG. 1

Therefore, as long as the marginal benefit from more public goods increases with the number of trips, the intramarginal term (13) must be positive.

Thus, when systematic preference differences are introduced into any particular variant of the model studied above, land rents will tend to undercapitalize the associated benefit. Operationally, this means that if land rents were used as a decision criterion, we might end up rejecting projects which ought to be accepted.

E. Sufficient Conditions for No External Capitalization

Having explored the conditions for internal capitalization in a community that was insulated from immigration forces, we turn now to a study of those external forces. We will show first that any systematic differences between towns can serve to neutralize the external forces at least with respect to their effect on small projects. There is an important caveat here, however. What we show is that communities could arrange things so as to insulate themselves; we cannot show that they have an incentive to follow such policies. Indeed, we have shown in some related work that communities do not have a myopic incentive to insulate in many instances.[3] We will ignore this issue here.

We employ the simplest model of community here, in which there are no choices with regard to public goods use. The reader might usefully recall that this is one of the cases in which there is no internal capitalization. All agents within the town are assumed alike. To make the point about potential insulation in the strongest possible way, we will also assume that preferences of all individuals in all towns are alike and that towns differ only in income levels; that difference alone

[3] See Starrett 1977.

is enough to insulate towns. Actually, this case may not be so unrealistic, since sorting by income level is the clearest observed Tiebout differentiation among communities.

Referring back to the first model in Section IIA, we can specify the characteristics of community k completely by a level of public goods (q^k), an income level of residents ($\bar{I}^k$),[4] and a cost level ($h^k = T^k + f^k[\bar{g}^k, s] + r_s^k \hat{\ell}_s$) which must, at equilibrium, be independent of location (s). It is convenient to assume that the $\bar{g}$ and $\hat{\ell}$ levels are the same everywhere so that they can be deleted from the analysis, although this is not really necessary. Note that, in the model as now constituted, there is no economic distinction between property taxation and lump-sum taxation. Now, the welfare (U^{kk}) of an individual with income $\bar{I}^k$ living in his own community can be expressed as $U^{kk} \equiv U(q^k, \bar{I}^k - h^k)$.

We now ask the following question. Suppose that a set of q's, I's, and h's is specified. Will anyone have an incentive to move immediately, and if not would anyone see such an incentive if one of the q's were changed marginally? We will show that an appropriate initial choice will make the answer to both these questions "no" as long as income effects are not negligible. Suppose a k agent considers moving to j. We assume that, if he does this, he must pay the taxes (and other costs) appropriate to j. (If, instead, he expects to pay the taxes appropriate to k, as he surely would if the tax were a property tax, the results are still the same, but we would have to incorporate the taxes in the I term rather than the h term.) On the other hand, an agent's income is determined by his place of initial ownership. This must be true of property income (as long as there are no unanticipated capital gains), and labor income is the same everywhere, by assumption. Hence, he would expect to get utility $U^{kj} \equiv U(q^j, I^k - h^j)$.

The move is not desirable, and would not be desirable for any sufficiently small variation in q^j, as long as $U^{kk} > U^{kj}$. Similarly, a move from j to k is not desirable as long as $U^{jk} \equiv U(q^k, \bar{I}^j - h^k) < U(q^j, \bar{I}^j - h^j) \equiv U^{jj}$. Are these two conditions consistent? Let us order j and k in such a way that $q^k > q^j$. Then stability certainly requires $h^k > h^j$. To see what else is required, it is convenient to change variables slightly. If we define $\Delta q = q^j - q^k$, $z^k = \bar{I}^k - h^k$, $z^j = \bar{I}^j - h^k$, and $\Delta z = h^k - h^j$, then the two stability conditions may be written as $U(q^k, z^k) > U(q^k + \Delta q, z^k + \Delta z)$ and $U(q^k, z^j) < U(q^k + \Delta q, z^j + \Delta z)$, with $\Delta q < 0$ and $\Delta z > 0$. These conditions are consistent as long as $\bar{I}^k > \bar{I}^j$, as can be seen in figure 2. Of course, we have assumed in the construction that the income effect is normal and significant, so that increases in income increase the marginal rate of substitution of z for q. Clearly, community k could

[4] The $\bar{I}^k$ is clearly the average income in the community.

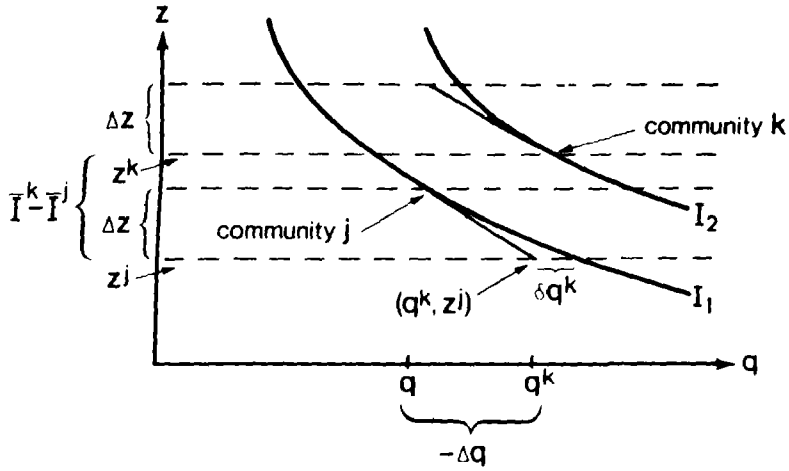


FIG. 2

engage in any project of size less than or equal to δq^k without inducing entry, while a similar statement holds for community j.

Naturally, if projects undertaken are too large, entry may be induced. However, we assert that any systematic differences between towns (direct preference differences would have done as well) can imply no external capitalization of marginal projects. This same threshold effect would result if the segregation of communities were caused by a fixed cost of moving (rather than by preference differences).[5]

It is interesting to note that if we now reintroduce the possibility of internal capitalization, its presence may be enough to neutralize external forces, even if there are no intrinsic differences at all among people. That is, if $q^k > q^j$ implies $R^k > R^j$, we will automatically have $\bar{I}^k > \bar{I}^j$, assuming that there are no other income differences.

We chose income differences to illustrate the possibility of insulation because they were the most innocuous-sounding differences we could think of. Obviously, preference differences are more likely to serve this purpose.[6] Indeed, some might argue that income effects are unimportant in the context discussed here. Suppose that we reexamine our problem in the absence of income effects. Interestingly, the whole character of the analysis changes.

[5] Of course, if there are no costs of moving and a continuum of preference types, any project will induce some entry and, presumably, some capitalization.

[6] Since the Tiebout model is characterized by preference difference among communities, there will be no external capitalization of small projects in Tiebout communities that are "sufficiently" differentiated.

F. Sufficient Conditions for Full External Capitalization

Without income effects or preference differences between communities it is impossible to insulate, and the resulting migration forces will generate external capitalization. To show this, let us reintroduce preference differences within communities while at the same time eliminating any income effects or preference differences between communities. An a-type resident of income class i now has preference of the form $\hat{V}^{ai} = \hat{U}^a(q) + I^i - h$. At this point, it is convenient to define a fictitious "average" resident of community k. This average resident clearly would have preference of the form $\bar{V}^k = \hat{\bar{U}}^k(q^k) + \bar{I}^k - h^k$, where $\bar{I}^k$ is the average income level in k (as before) and $\hat{\bar{U}}^k$ is the average felicity function in k. Our assumption of no preference differences between communities will be taken to mean that the average felicity functions are the same in all communities, that is, $\hat{\bar{U}}^k(q) = \hat{\bar{U}}(q)$, all k.

Now let us reconsider the stability conditions for equilibrium. Clearly, a necessary condition for stability is that our fictitious average resident does not want to move in either direction between any pair of communities. The corresponding pair of inequalities for communities k and j are

$$\hat{\bar{U}}(q^k) + \bar{I}^k - h^k \geq \hat{\bar{U}}(q^j) + \bar{I}^k - h^j \tag{14}$$

and

$$\hat{\bar{U}}(q^k) + \bar{I}^j - h^k \leq \hat{\bar{U}}(q^j) + \bar{I}^j - h^j. \tag{15}$$

Clearly strict inequality in either direction is impossible, and both relations must hold as equalities in any equilibrium situation; the average resident must be indifferent at the margin as to where to live. It follows that any welfare-improving project anywhere must induce entry.

We can now use equations (14) and (15) to generate a simple external capitalization result. These equations must hold both before and after a project (in some particular town z), so we can differentiate them with respect to q^z. In doing so we must allow for the possibility that a change in q^z could affect I^j and h^j ($j \neq z$). We have already seen one way in which this could happen through external ownership of local firms; now that migration is certain, there are the so-called fiscal externalities as well; when one community attracts residents from other communities, the lost tax revenue is a real external cost to the "losing" communities. For a rigorous demonstration of this fact, the reader is referred to Starrett (1980).

Allowing for potential changes in all variables, we can write

$$dV^z \equiv d\hat{\bar{U}}(q^z) + d\bar{I}^z - dh^z = d\bar{I}^z - dh^j \tag{16}$$

and

$$dV^j = d\hat{U}(q^z) + d\bar{I}^j - dh^z = d\bar{I}^j - dh^j, \tag{17}$$

where dV^i is shorthand notation for the change in average welfare of community i.

Let us define $dW^\#$ to be the total external effect; that is,

$$dW^\# = \sum_j N^j dV^j = dI^\# - \sum_{j \neq z} N^j dh^j.^7$$

And $dW^z = N^z dV^z$ will stand for the welfare change in community z. Clearly, $dW = dW^\# + dW^z$. Simple substitutions yield the following two equations for the change in community z welfare and the change in social welfare:

$$dW^z = \frac{N^z}{N^\#} dW^\# - \frac{N^z}{N^\#}(dR^\# + d\Pi^\#) + dR^z + d\Pi^z$$

and

$$dW = \frac{N}{N^\#} dW^\# - \frac{N^z}{N^\#}(dR^\# + d\Pi^\#) + dR^z + d\Pi^z.$$

Finally, we must distinguish cases again, according to the nature of firm ownership.

1. Neutral Ownership

Suppose that each community owns its share of national profits. Then $N^z\Pi^\# - N^\#\Pi^z = 0$, and we have

$$dW^z = \frac{N^z}{N^\#} dW^\# + dR^z - \frac{N^z}{N^\#} dR^\# \tag{18}$$

and

$$dW = \frac{N}{N^\#} dW^\# + dR^z - \frac{N}{N^\#} dR^\#. \tag{19}$$

Clearly, total rent change in z is a good approximation to net welfare benefits in z as long as community z is small relative to the nation (so that $N^z/N^\#$ is small). However, net social benefits will be misrepresented by rent increases to the extent that the external effect is significant.

How important is this misrepresentation? It turns out it may be

[7] The I^i stands for total income in community i; the symbol # means the associated variable is aggregated over all communities except z. Recall that we are assuming that all local public goods projects are paid for locally. Naturally we could correct for the externalities if we could impose the optimal revenue-sharing scheme.

quite important even if $N^z/N^\#$ is small. We have studied the measurement of $dW^\#$ elsewhere (see Starrett 1980) and merely report the main results here.

When community z engages in a project, that project is certain to induce some entry. Other communities suffer real welfare losses which are equal to lost tax revenue (from the migrators) minus any attendant improvement in congestion "costs." Unless communities have already seriously overexpanded, the net benefit to outsiders is negative ($dW^\# < 0$). Hence even changes in residential land values will overstate net social benefit, quite generally.

Of course, if community z is not small as a fraction of the country, the theory takes a somewhat more complicated general form. Aside from the external terms, net benefits are capitalized into the difference between the change in land value in z and the "average" change in land values elsewhere. This formulation is reminiscent of that presented in Strotz (1968) for two regions.

There is one other major difference from the internal capitalization results: net benefits are externally capitalized, while gross benefits were internally capitalized. The economic reason for this discrepancy is not difficult to find. The marginal potential resident pays local taxes in the internal model (he is the boundary resident), while the marginal potential resident does not pay local taxes in the external model.

2. Local Ownership

If each community owns its own firms, then $d\Pi^\# = -dR_x^\#$, and we have

$$dW^z = \frac{N^z}{N^\#} dW^\# + dR_r^z - \frac{N^z}{N^\#} dR_r^\# \tag{20}$$

and

$$dW = \frac{N}{N^\#} dW^\# + dR_r^z - \frac{N^z}{N^\#} dR_r^\#. \tag{21}$$

The form of the results is the same except that capitalization is into residential rather than total land value. The reader might think at first that there is a contradiction between equations (19) and (21). After all, it cannot matter from a social point of view where the profits are owned, so why do these equations look different? The answer is that any increased cost of local industrial land will be reflected in the term $dW^\#$ in equation (19) but not in (21). Thus, even in the case of national ownership, only increases in residential land values represent real social benefit, once adverse land value externalities are canceled out.

III. Summary

We have shown in the previous sections that there is considerable diversity in the form and extent of capitalization, depending on the nature of the assumptions made. However, some fairly general principles emerge, and we summarize these in this section.

First, all of the models studied lead to the conclusion that, from the point of view of society as a whole, it is residential property values that capitalize project benefits (rather than total land value). The theories differ on the appropriate land base only from the perspective of the project-generating community; if lost profits from increased land values are exported, then increases in total land value measure benefits to the community, but, if not, then it is residential land values.

The internal and external capitalization models generally disagree on whether gross or net benefits are capitalized. However, it is worth pointing out one important case in which the two theories agree on this issue. This is the case of property taxation with agricultural zoning, for which both theories suggest that increases in net-of-tax land values capitalize net welfare benefits (and, naturally, increases in gross-of-tax land values capitalize gross benefits).

The main differences between internal and external capitalization results can be traced to the boundary conditions imposed on a representative community. Notice that such boundary conditions play no role at all in external capitalization; indeed, the external results hold regardless of what happens to rents or benefits at the boundary of the communities.

However, internal capitalization results vary considerably, depending on the assumptions governing behavior of land rent and the impact of incremental projects, at the boundary. In the extreme case in which the impact is nil and the (net-of-tax) rent goes down by the rate of taxation, internal capitalization agrees with external capitalization. But if boundary rent is unchanged (as it will be under many plausible conditions) the gross project benefits are internally capitalized. And if marginal projects actually make boundary residents worse off (because of a dominating congestion effect) then land values will overcapitalize even gross benefits.

We should recognize that, in all cases considered, land values will fail to capture some of the external benefits (or, more likely, costs) of a project. When the only external effect is through nationally owned profits, as was the case in the pure internal capitalization model, we could correct for the externality by using residential land value rather than total land value. However, in the free-trade, free-migration model of external capitalization, the external costs of a project are

generally much larger, and we conclude that increases in residential land value will overstate true net social benefit by an amount closely related to the size of the fiscal externality.

We close with some remarks on the basic assumptions underlying capitalization results. The two types of capitalization rest on different, and to some extent complementary, assumptions. External capitalization requires that the distribution of agents by economic characteristics be the same in one community as in all others, while internal capitalization requires that agents within a particular community must have similar economic characteristics (though these characteristics may differ from community to community). Thus, the internal capitalization results are most applicable in a "pure Tiebout" world where each community specializes and attracts a single type of agent. The external capitalization model is most applicable in the opposite-extreme world of "pure scrambling" in which all communities look alike in terms of the characteristics of agents. Naturally, there are many intermediate cases between these extremes (including the real world, presumably!). To the extent that neither extreme is a good approximation, capitalization will not hold in either form, or will hold only partially.

References

Freeman, A. Myrick, III. "Air Pollution and Property Value: A Methodological Comment." *Rev. Econ. and Statis.* 53 (November 1971): 415–16.

Hamilton, Bruce W. "Property Taxes and the Tiebout Hypothesis: Some Empirical Evidence." In *Fiscal Zoning and Land Use Controls: The Economic Issues*, edited by Edwin S. Mills and Wallace E. Oates. Lexington, Mass.: Lexington, 1975.

Lind, Robert C. "Spatial Equilibrium, the Theory of Rents, and the Measurement of Benefits from Public Programs." *Q.J.E.* 87 (May 1973): 188–207.

Oates, Wallace E. "The Effects of Property Taxes and Local Public Spending on Property Values: An Empirical Study of Tax Capitalization and the Tiebout Hypothesis." *J.P.E.* 77, no. 6 (November/December 1969): 957–71.

Pines, David, and Weiss, Yoram. "Land Improvement Projects and Land Values." *J. Urban Econ.* 3 (January 1976): 1–13.

Polinsky, A. Mitchell, and Rubinfeld, Daniel L. "The Long Run Effects of a Residential Property Tax and Local Public Services." *J. Urban Econ.* 5 (April 1978): 241–62.

Polinsky, A. Mitchell, and Shavell, Steven. "Amenities and Property Values in a Model of an Urban Area." *J. Public Econ.* 5 (January/February 1976): 119–29.

Starrett, David A. "On the Capitalization Hypothesis in Closed Communities." Technical Report no. 241. Inst. Math. Studies Soc. Sci., Stanford Univ., 1977.

———. "On the Capitalization Hypothesis for Local Public Finance." Unpublished manuscript, Stanford Univ., June 1978.

———. "On the Method of Taxation and the Provision of Local Public Goods." *A.E.R.* 70 (June 1980): 380–92.

Strotz, Robert H. "The Use of Land Rent Charges to Measure the Welfare Benefits of Land Improvement." In *The New Economics of Regulated Industries: Rate-Making in a Dynamic Economy*, edited by Joseph E. Haring. Econ. Res. Center, Occidental College, Los Angeles, 1968.

Tiebout, Charles M. "A Pure Theory of Local Expenditures." *J.P.E.* 64, no. 5 (October 1956): 416–24.

Money and the Dispersion of Relative Prices

Zvi Hercowitz

University of Rochester and National Bureau of Economic Research

A price dispersion equation is tested with data from the German hyperinflation. The equation is derived from a version of Lucas's and Barro's partial information-localized market models. In this extension, different excess demand elasticities across commodities imply a testable dispersion equation, in which the explanatory variable is the magnitude of the unperceived money growth. In order to test this hypothesis a price dispersion series is constructed, and a measure of the unperceived part of money growth is estimated. The model receives support from the empirical analysis, although it is evident that unincluded variables have important effects on price dispersion

The existence of a positive correlation between absolute price level variability and the dispersion of relative prices was observed by, for example, Mills (1927), Graham (1930), and recently by Vining and Elwertowski (1976). In his study of U.S. price behavior during the period 1920–26, Mills (1927) says: "We have not however exhausted the possibility of discovering a relationship between price level and dispersion. It may be that dispersion depends upon the violence of the price change, regardless of direction" (p. 284). Graham (1930) finds in the post–World War I German hyperinflation an additional dynamic element of price behavior: "It is clear that with the initiation of an upward movement in general prices a series of lags in individual

This paper is part of my doctoral dissertation at the University of Rochester. I wish to express my gratitude to Robert Barro for his guidance and helpful suggestions I have also benefited from discussions with Bill Bomberger, Rudiger Dornbusch, Chip Miller, Mark Rush, Hal White, and from insightful comments by a referee of this journal. Financial support from the National Science Foundation is gratefully acknowledged.

Journal of Political Economy, 1981, vol. 89, no 2)
© 1981 by The University of Chicago 0022-3808/81/8902-0004$01 50

prices developed, that these lags tended quickly to disappear when stability of general prices was reached on a new level, or when general prices fell, but that they were nevertheless progressively eliminated even though the general price level continued to rise" (p. 175). This observation suggests that unexpected events may have an important role in the determination of price dispersion. Individual prices disperse at the beginning of an upward swing in the price level when the acceleration is presumably unexpected. As inflation continues the element of surprise wanes, and prices tend to converge.

The studies cited above failed to offer an economic rationale for the observed statistical correlation. Recently, a theoretical explanation of the relationship between price level variance and relative price dispersion was offered by Barro (1976). Using a localized markets framework of the type described by Phelps (1970) and employed by Lucas (1973), Barro links the dispersion of relative prices to the variance of the money supply. The key elements of this model are, on the one hand, individuals possessing incomplete current information and, on the other, demand and supply in each market reacting to relative prices as they are locally perceived. Thus, agents are confronted with the problem of determining whether locally observed price movements are caused by general inflation or by shifts in relative excess demand. The larger the variance of the money supply, the more likely are agents to attribute local price movements to general inflation rather than to relative shifts. Accordingly, as the money variance rises, local price changes induce smaller supply-and-demand responses—that is, excess demand becomes less elastic. Consequently, stochastic shifts to local excess demand produce larger changes in individual prices, so that the dispersion of prices across markets tends to increase with the variance of money. In this specification of the model, in which all markets have the same structure, dispersion is unrelated to the magnitude of realized money shocks.

This paper modifies Barro's framework by interpreting each location to be the market of a specific commodity, characterized by a particular excess demand elasticity. Because elasticities vary across markets, aggregate shocks affect each commodity price differently. Therefore, in this modified setup price dispersion is positively related to the magnitude of these shocks.

The model also predicts that systematic or perceived money growth is neutral with respect to price relationships. Accordingly, a money shock in this model is defined to be the component of money growth that is currently unobservable and cannot be inferred from currently available information. Whereas the quotation from Graham suggested that sudden—presumably unexpected—shifts in money growth cause dispersion, in this model unexpected monetary expan-

sion disperses prices only if it is, at least partially, currently unperceived.

The main task of this study is to evaluate this hypothesis with data from the German hyperinflation, a period of predominantly monetary disturbances. The period considered runs from January 1921 to July 1923. The vertiginous monetary expansion initiated in August 1923 differentiates the last phase of the hyperinflation, and, thus, it is not included in the sample.

The theoretical framework, presented in Section I, neglects some important facets of the hyperinflation. Specifically, it ignores the foreign exchange market and the sustained divergence between the internal and external values of the mark,[1] which obviously are related to relative prices. Also ignored in the main text are changes in the velocity of monetary circulation. However, in a brief discussion some general conditions are given under which price dispersion is neutral with respect to velocity changes. Finally, the only aggregate exogenous disturbances assumed to be affecting the economy are periodic infusions of new money made by the government. Real aggregate shocks are ignored. Empirically, they are probably of relatively minor importance and can be considered to be part of the error term in the estimated equations.

The testing of the dispersion equation requires two important preliminary steps. First, a price dispersion series is computed in Section II using an interesting set of data. It consists of monthly averages of 68 commodity prices, ranging from foods to metals, for the period of January 1921 through July 1923 (31 months). Data were unavailable for the months prior to January 1921. Next, a money growth equation is set up and estimated in Section III. To do so, both an information set available to agents economy-wide and a functional form relating this set to money creation are postulated and discussed. The explained part of money growth in the estimated equation is taken as a measure of the perceived rate of monetary expansion. Correspondingly, the unexplained part is interpreted as the money growth rate that could not be perceived from the assumed information set. These figures, in conjunction with the price dispersion series, are used in Section IV to test the price dispersion equation.

Parks (1978) has also tested a model of price dispersion using pre- and post-World War II U.S. data. In Parks's model dispersion is explained by changes in real income and the unexpected part of inflation, as measured by the innovation in the inflation rate. In this specification expected inflation and changes in real income are treated as exogenous variables. He finds a strong positive correlation

[1] See, e.g., Bresciani-Turroni 1937.

between unexpected inflation and price dispersion. His tests also suggest a separate but smaller effect of the actual inflation rate.

The present paper estimates an equation that relates price change dispersion to the exogenous shocks affecting the economy, in this case unperceived monetary injections. An additional monetary variable that is theoretically relevant for price dispersion is the variance of money shocks. An estimate for this variance is obtained from the money growth analysis and is included in the estimation. The model receives significant support from the empirical analysis. In particular, the variable measuring unperceived money growth has substantial explanatory power for price dispersion. The results also make clear that unincluded variables have important effects on price dispersion. Some of these are briefly considered in Section V.

I. The Model

The economy consists of an arbitrarily large number of physically separated markets indexed by z. In each location a specific commodity is produced and traded. At each date t the agents, assumed to be risk neutral, exchange money only for the commodity being traded in the market in which they are currently located. At date $t + 1$, agents change location at random, and the process is repeated. Consider now the information set available to the agents. It contains not only lagged values of all relevant variables, but also current information which is limited to the local market price $P_t(z)$, and some economy-wide shared knowledge about current variables related to money creation. Actual money growth, however, includes a random term which is assumed unknown.

The supply and demand for commodity z assume the log-linear forms:

$$y_t^s(z) = \alpha^s(z)[P_t(z) - EP_t] + \epsilon_t^s(z), \tag{1}$$

$$y_t^d(z) = -\alpha^d[P_t(z) - EP_t] + (M_t - EP_t) + \epsilon_t^d(z), \alpha^s(z) > 0, \alpha^d > 0. \tag{2}$$

The operator E is the mathematical expectation taken conditional on the information available in market z at time t. For each commodity z, $P_t(z) - EP_t$ is the locally perceived relative price. The expressions $\epsilon_t^s(z)$ and $\epsilon_t^d(z)$ represent relative shifts to supply and demand, respectively. The excess demand shift, $\epsilon_t(z) \equiv \epsilon_t^d(z) - \epsilon_t^s(z)$, is assumed serially uncorrelated, normally distributed with zero mean and variance σ_ϵ^2. This variance is assumed to be equal in all markets. For each z, $\alpha^s(z)$ is the short-run relative price elasticity of supply. Disparity in the supply elasticities of different goods follows from heterogeneous production functions. However, in the long run relative prices are assumed

fixed because of perfect substitutability on the supply side. The long run is measured here by one period, after which all suppliers can shift to other markets.

Looking one period ahead, all the markets offer the same mean price, but, as shown below, the corresponding variances differ according to the excess demand elasticities. Because agents are risk neutral, they are indifferent between the markets, and thus they choose a market for the next period randomly. There is an additional point related to the ex ante variability of the individual prices. Intuitively, one would expect a market with more price variance to be less desirable because local information would yield price level estimates of lower precision. However, as shown below, this turns out not to be the case.

On the demand side, the relative price elasticities are assumed constant across markets. The demand function also includes the term $M_t - EP_t$, which accounts for a real balance effect.

At the beginning of each period the stock of money in the economy is increased by transfers from the government to the public. This new money is assumed to be distributed equally across the markets. Within each market, however, the transfers are allocated randomly among a large number of agents. The rate of growth of the money stock, $m_t = M_t - M_{t-1}$, obeys

$$m_t = \sum_i \beta_i X_{it} + \tilde{m}_t \equiv g_t + \tilde{m}_t,$$

where the X_{it}'s are variables (past or current) that can be observed in all locations and the β's are known coefficients. The quantity $\tilde{m}_t$ is a random variable with zero mean and variance σ_m^2. Thus, g_t is the expectation about money growth formed from all the economy-wide shared information. It can be considered the prior expectation. The posterior is formed using the additional information conveyed by the local price. Thus, while g_t is the same everywhere, the posterior expectation Em_t is conditional on location-specific information as well and therefore varies across markets.

From equations (1) and (2) market clearing implies that

$$P_t(z) = \{1 - 1/[\alpha^s(z) + \alpha^d]\} EP_t + \{1/[\alpha^s(z) + \alpha^d]\} [M_t + \epsilon_t(z)]. \quad (3)$$

For each z, the sum $\alpha^s(z) + \alpha^d$ is the relative price elasticity of excess demand. Let $\lambda(z) \equiv 1/[\alpha^s(z) + \alpha^d]$. Each market has a constant $\lambda(z)$, but across markets $\lambda(z)$ is distributed according to a given density function with average value λ and "variance" σ_λ^2. Consistent with the assumption that agents possess accurate knowledge about the structure of the economy, this distribution is assumed to be known.

Following Lucas (1973) and Barro (1976), the solution for prices in terms of exogenous variables is obtained using the method of unde-

termined coefficients. Given the log-linearity of the model, the solution for the aggregate price level has the form

$$P_t = \Pi_1 M_{t-1} + \Pi_2 g_t + \Pi_3 \tilde{m}_t. \tag{4}$$

Namely, the aggregate price level will be related to the current money stock, which is divided into its different components. Lagged values, if added to (4), yield zero coefficients. Since M_{t-1} and g_t are fully perceived at time t, taking the expectation of both sides yields

$$EP_t = \Pi_1 M_{t-1} + \Pi_2 g_t + \Pi_3 E\tilde{m}_t. \tag{5}$$

The conditional expectation of m_t is now computed. Rewrite (3) as:

$$\delta_t(z) = g_t + \tilde{m}_t + \epsilon_t(z),$$

where

$$\delta_t(z) = [1/\lambda(z)]P_t(z) - [1/\lambda(z) - 1]EP_t - M_{t-1}.$$

The total disturbance affecting market z, $\delta(z)$, is partly nominal and partly real. Agents perceive $\delta(z)$ and form their expectations about its components. Given the stochastic specification of m_t and $\epsilon_t(z)$, the mean of the distribution of m_t conditional on $\delta(z)$ is

$$Em_t = g_t + \frac{\sigma_m^2}{\sigma_m^2 + \sigma_\epsilon^2}[\delta_t(z) - g_t],$$

or

$$Em_t = g_t + \frac{\sigma_m^2}{\sigma_m^2 + \sigma_\epsilon^2}[\tilde{m}_t + \epsilon_t(z)]. \tag{6}$$

Observe that $\lambda(z)$ does not appear in (6). Since agents located in z know this elasticity, they are able to isolate the composite disturbance independently of $\lambda(z)$. Thus, while the ex ante variance of prices depends on the particular elasticity (this follows from eq. [26] below), the precision obtainable from the local information is independent of $\lambda(z)$. Indeed $P_t(z)$ would convey less valuable information in higher price variance markets if the differential variability was due to a disparity in σ_ϵ^2. In this model, however, σ_ϵ^2 is the same across markets.

Substitute now (6) into (5) and the resulting expression for EP_t into (3) to obtain

$$EP_t = \Pi_1 M_{t-1} + \Pi_2 g_t + \Pi_3 \frac{\sigma_m^2}{\sigma_m^2 + \sigma_\epsilon^2}[\tilde{m}_t + \epsilon_t(z)], \tag{7}$$

$$\begin{aligned} P_t(z) = {} & [1 - \lambda(z)]\left\{\Pi_1 M_{t-1} + \Pi_2 g_t + \Pi_3 \frac{\sigma_m^2}{\sigma_m^2 + \sigma_\epsilon^2}[\tilde{m}_t + \epsilon_t(z)]\right\} \\ & + \lambda(z)[M_{t-1} + g_t + \tilde{m}_t + \epsilon_t(z)]. \end{aligned} \tag{8}$$

A new expression for the general price level can be computed from (8) by averaging with respect to the densities of $\lambda(z)$ and $\epsilon_t(z)$:

$$P_t = (1 - \lambda)\left(\Pi_1 M_{t-1} + \Pi_2 g_t + \Pi_3 \frac{\sigma_m^2}{\sigma_m^2 + \sigma_\epsilon^2}\tilde{m}_t\right) + \lambda(M_{t-1} + g_t + \tilde{m}_t). \tag{9}$$

Since equation (9) is identical to (4), the solution for Π_1, Π_2, and Π_3 is obtained by equating the corresponding coefficients in the two equations:

$$\begin{aligned} \Pi_1 &= 1, \\ \Pi_2 &= 1, \\ \Pi_3 &= \frac{\sigma_m^2 + \sigma_\epsilon^2}{\sigma_m^2 + (1/\lambda)\sigma_\epsilon^2}. \end{aligned} \tag{10}$$

Substituting (10) into (8) and (9) and rearranging terms yields the solution for the individual commodity price and the average price level:

$$P_t(z) = M_{t-1} + g_t + \frac{\sigma_m^2 + \lambda(z)(1/\lambda)\sigma_\epsilon^2}{\sigma_m^2 + (1/\lambda)\sigma_\epsilon^2}[\tilde{m}_t + \epsilon_t(z)], \tag{11}$$

$$P_t = M_{t-1} + g_t + \frac{\sigma_m^2 + \sigma_\epsilon^2}{\sigma_m^2 + (1/\lambda)\sigma_\epsilon^2}\tilde{m}_t. \tag{12}$$

The resulting actual relative price is:

$$P_t(z) - P_t = (1 - \theta)\tilde{\lambda}(z)\tilde{m}_t + [\theta + \lambda(z)(1 - \theta)]\epsilon_t(z), \tag{13}$$

where $\tilde{\lambda}(z) \equiv \lambda(z) - \lambda$ and $\theta \equiv \sigma_m^2/[\sigma_m^2 + (1/\lambda)\sigma_\epsilon^2]$.

The hypothesis expressed by equation (13) is that only the unperceived part of money growth can affect price relationships. Note that the realized values of the unperceived money growth appear in the relative price expression. This follows from the confusion between $\tilde{m}_t$ and $\epsilon_t(z)$. Since in general $E\tilde{m}_t \neq \tilde{m}_t$, part of the money shocks is mistakenly perceived to be a shift in relative excess demand. The ensuing short-run supply reactions differ across markets according to $\alpha^s(z)$, thus causing dispersion among actual prices. On the other hand, g_t is correctly identified as an aggregate disturbance and therefore cannot be confused with a relative shift of excess demand. The neutrality of perceived money follows from $y_t^s(z)$ and $y_t^d(z)$ being functions of the relative price and from the one-to-one relationship between g_t and the expected price level (eqq. [7] and [10]). Given some value for g_t, the quantities along the supply-and-demand schedules are the same as before, for local nominal prices higher by an amount equal to

the adjustment in EP_t—which equals g_t. Therefore, the market clears at a $P_t(z)$ which is higher by the same degree in all markets.[2]

The variance of relative prices at time t, defined as $\tau_t^2 \equiv 1/N \sum_{z=1}^{N} [P_t(z) - P_t]^2$, where N is the "very large" total number of markets in the economy, can be computed from equation (13)[3]

$$\tau_t^2 = \{(1 - \theta)^2\sigma_\lambda^2 + [\theta + \lambda(1 - \theta)]^2\}\sigma_\epsilon^2 + (1 - \theta)^2\sigma_\lambda^2\bar{m}_t^2. \qquad (14)$$

An empirical test of this equation requires a measure of dispersion among prices or price indexes of different commodities. Mills (1927, chap. 3) discusses problems that the interpretation of this dispersion measure presents. For example, long-run differential technological changes will cause prices to disperse over time. One would like to filter out such effects, because the focus here is on short-run distortions caused by incomplete current information. The problem is alleviated by using rates of price *change* rather than price levels. Different trends do not affect the variation of price change dispersion over time—although alterations in these trends will. Thus, some of the long-run relative prive movements effects can be filtered from the dispersion measure. What remains can be considered to be captured by the random term in the dispersion equation.

The variance of the rates of change in individual prices is calculated using equation (13) and the equivalent for $t - 1$. This variance, defined as

$$\gamma_t^2 \equiv \frac{1}{N}\sum_{z=1}^{N} \{[P_t(z) - P_{t-1}(z)] - (P_t - P_{t-1})\}^2, \qquad (15)$$

follows as

$$\gamma_t^2 = 2(1 - \theta)^2\sigma_\lambda^2\sigma_\epsilon^2 + 2[\theta + \lambda(1 - \theta)^2]\sigma_\epsilon^2 + (1 - \theta)^2\sigma_\lambda^2(\bar{m}_t - \bar{m}_{t-1})^2.$$

Equation (15) is the final price dispersion equation that is generated

[2] In order to consider the effects of changes in the velocity of money circulation on price dispersion, I worked out a similar simple model in which there is some current public information about future money growth. This information may be conveyed by political or military events that are believed to have implications for the future state of government finances. The prediction of future monetary expansion, which generates inflationary expectations, affects the velocity of circulation in the current period. With respect to relative prices, if the knowledge about future money growth is shared economy-wide, they will be unaffected by the change in velocity. This neutral effect follows from the same mechanism determining the neutrality of perceived money. Since all agents share the same knowledge and are assumed to use it in the same model to predict its effects, they will equally adjust their EP_t according to the change in velocity taking place. A "one-time jump" in all prices therefore occurs, without affecting their dispersion.

[3] For this computation, since $\epsilon_t(z)$ and $[\epsilon_t(z)]^2$ are independent of $\lambda(z)$, $[\lambda(z)]^2$, and $\bar{\lambda}(z)$, the following equalities are used: $(1/N)\Sigma\epsilon_t(z)\cdot\lambda(z) = 0\cdot\lambda = 0$, $(1/N)\Sigma[\epsilon_t(z)]^2[\bar{\lambda}(z)]^2 = \sigma_\epsilon^2\sigma_\lambda^2$, and $(1/N)\Sigma[\epsilon_t(z)]^2[\lambda(z)]^2 = \sigma_\epsilon^2(\sigma_\lambda^2 + \lambda^2)$.

by the model. Because it deals with the dispersion of price changes, the appropriate monetary-shocks variable is the magnitude of changes in $\tilde{m}_t$.

Consider next the implied relationship between the variance of money shocks and the dispersion of relative prices. Barro's theoretical result was that σ_m^2 is positively correlated with relative price variability. However, the effect of σ_m^2 is ambiguous in this extended version of the model, since it has different and opposite effects on the three terms in the γ_t^2 expression. The second term on the right-hand side of (15) is the remainder of the expression when all markets are alike; that is, when, as in Barro's case, all have the same excess demand elasticity ($\sigma_\lambda^2 = 0$). This term corresponds to his relative price variance, which depends positively on σ_m^2 when $0 < \lambda < 1$. This condition is the counterpart to Barro's assumption that substitution effects dominate wealth effects.

The first term accounts for the positive interaction between the diversity in elasticities and the strength of the relative shifts. A term of this sort would be included also in the dispersion expression under full current information. In the present case of partial information, the fraction $(1 - \theta)$ appears here because agents typically underestimate the magnitude of the relative shifts, thus diminishing their effect on price dispersion. Because this underestimation increases with σ_m^2, the first term is negatively related to the money variance. The other, more interesting negative effect of σ_m^2 appears in the third term, namely, in the coefficient of $(\tilde{m}_t - \tilde{m}_{t-1})^2$. If σ_m^2 increases—or more precisely when the public perceives it doing so—money disturbances are less confused with real shifts, implying that a given shock induces smaller dispersion. This effect is a relative price equivalent of Lucas's hypothesis about the link between the variance of the nominal disturbances and the slope of the Phillips curve.

In the testing of equation (15), reported in Section IV, an attempt is made to capture the different effects of σ_m^2 and its net influence on price dispersion. However, the procedure adopted does not indicate that shifts in σ_m^2 have an important effect.

II. Construction of the Price Dispersion Series

This section reports the computation of a measure of price dispersion for the hyperinflation in Germany during the period January 1921–July 1923. The data set, consisting of 68 series of monthly averages of wholesale commodity prices, is obtained from the German statistical yearbook issues of 1921/22 and 1923 (see *Statistisches Reichsamt* 1921/22 and 1923). Other series from this source, some reported only until December 1921 (seven commodities) and others beginning only in

January 1922 (21 commodities), were deleted in order not to introduce a bias due to changes in the sample size and composition.

Prices are quoted from commodity exchanges of several German cities.[4] Each series, however, originates in a single location. The 68 commodities include 27 foodstuffs, 19 textiles and leathers, and 22 metals, oils, and coals. They are not finished goods but materials in a rather raw state. Because weights for the different commodities are unfortunately not available, unweighted rates of price change are used. Hopefully, the wide range of commodities in the sample approximates the general relative price instability during that period.

The individual price rates of change are computed as the first difference of the logarithms of the prices. Average values and variances are then calculated using

$$\Delta P_t = \frac{1}{N} \sum_i \Delta P_{it},$$

$$\gamma_t^2 = \frac{1}{N} \sum_i (\Delta P_{it})^2 - (\Delta P_t)^2,$$

where P_{it} is the price of commodity i and $\Delta P_{it} = \log P_{it} - \log P_{it-1}$. Table 1 contains the computed values of ΔP_t and γ_t^2. Due to missing observations, the actual number of commodities included in the calculations varies slightly from month to month. The third column in table 1 indicates the number of commodities for which both P_{it} and P_{it-1} are available.

III. Estimation of the Unperceived Part of Money Growth

Determination of the unperceived component of money growth during the hyperinflation requires a specification of the information set assumed to have been available to the public and the functional form for calculating the conditional expectation of money growth. Consider the expectation conditioned on economy-wide or "global" information g_t. This term was defined in Section I to be the prior expectation and is distinguished from the posterior expectation because it does not incorporate the additional information derived from local price observations.

This global information is assumed to consist of the current government spending in foreign exchange units, S_t, the current exchange rate, e_t, and 1-month lagged data on the money stock, price level, and all other macroeconomic variables. Not included is government reve-

[4] Given this source, these data do not present the problem of reported wholesale price data in the United States, discussed by Stigler and Kindahl (1970), that they do not always reflect discounts from list prices.

TABLE 1

MEAN AND VARIANCE OF WHOLESALE RATES OF PRICE CHANGES, GERMANY, FEBRUARY 1921–JULY 1923

Month	ΔP_t	γ_t^2	Number of Commodities
1921.			
February	−.09	.015	63
March	−.05	.011	66
April	−.02	.007	66
May	−.01	.026	66
June	.05	.032	67
July	.04	.034	66
August	.19	.081	66
September	.19	.033	68
October	.24	.034	68
November	.38	.043	68
December	−.03	.062	66
1922.			
January	.05	.018	66
February	.12	.013	68
March	.25	.011	65
April	.12	.018	65
May	.04	.013	66
June	.09	.007	66
July	.36	.017	67
August	.70	.097	68
September	.40	.071	66
October	.68	.064	66
November	.78	.038	66
December	.22	.052	66
1923:			
January	.70	.064	66
February	.69	.099	66
March	−.17	.041	65
April	.11	.018	65
May	.50	.039	66
June	.82	.048	62
July	1.25	.151	63

SOURCE.—Based on monthly average price data from *Statistisches Jahrbuch für das Deutsche Reich* (1921/22), pp. 282–83, and (1923), pp. 286–89.

nue from taxation and other sources because this variable depends on the current level of economic activity and is unlikely to be preannounced and to be widely known contemporaneously. It is natural to assume that the part of government expenditure consisting of the reparations to the Allied Powers was known in foreign exchange terms. With respect to the other expenditures, the implication is that nominal spending was observable and could be readily converted given the exchange rate.[5]

[5] A referee of this *Journal* suggested an alternative procedure for isolating unper-

The prior expectation of money growth is derived from the government monthly budget constraint, namely,

$$M_t^0 - M_{t-1}^0 = S_t^0 e_t^0 - (\text{other forms of nominal government revenue}). \tag{16}$$

The superscript 0 indicates that the variables are not in logs but in their original form. Equation (16) indicates that creation of high-powered money equals the part of nominal expenditure that is not financed in some other way. The expression M_t^0 would correspond here to the end-of-month money stock. The other forms of government finance are taxes, net sale of bills, gold sold to the public, etc.

If this other revenue comprised an approximately fixed proportion of total expenditure over time, the budget equation above could be expressed as

$$M_t^0 - M_{t-1}^0 = kS_t^0 e_t^0 + \text{random term}, \tag{17}$$

where k $(0 < k < 1)$ is the average fraction of the expenditure financed by money issue.

The first attempt to generate a perceived money growth series was made using an equation of this type. Dividing (17) through by M_{t-1}^0, money growth appears linearly related to $S_t^0 e_t^0 / M_{t-1}^0$. The three variables in this ratio are assumed currently known, and, therefore, this specification is consistent with the notion that the conditional expectation can be formed using only currently observable variables.

However, a regression of this form,[6] including a constant, shows that k was probably not constant over the period. Specifically, the existence and pattern of residual serial correlation,[7] plus some additional considerations discussed below, suggest a nonlinear relationship between money issue and spending during that period.

Assuming, then, that the fraction k is not constant over time, the question is whether something can be said about its determinants. In

ceived money using information from the Central Bank's balance sheet. This information is available monthly in Statistisches Reichsamt (1925). When the Central Bank discounted a bill, the corresponding credit appeared in the current account of the state or the private borrower. Only when these funds were withdrawn did the notes in circulation increase. Using information about these credits, agents could forecast money growth in the near future.

[6] The estimated OLS equation is the following (see below for the inclusion of the lagged spending variable):

$$\frac{M_t^0 - M_{t-1}^0}{M_{t-1}^0} = \underset{(.019)}{-.049} + \underset{(.033)}{.317}\,\frac{S_t^0 e_t^0}{M_{t-1}^0} + \underset{(.089)}{.381}\,\frac{S_{t-1}^0 e_{t-1}^0}{M_{t-1}^0}$$

$$(R^2 = .95,\ \text{D-W} = 1.0,\ \sigma = .067).$$

[7] The residuals are generally negative at the beginning and the end of the period, approximately the low and high values of money growth, and generally positive for the rest of the sample.

order to suggest an answer to this question, rewrite equation (17) as

$$\frac{M_t^0 - M_{t-1}^0}{M_t^0} = k_t \frac{S_t^0}{M_t^0/e_t^0} + \text{random term.} \qquad (18)$$

Equation (18) preserves the positive correlation between money growth and the ratio of real expenditure to real cash balances, but, unlike (17), the fraction k is now allowed to vary over time. It is now argued that k_t is itself correlated with M_t^0/e_t^0 and S_t^0.

To examine this correlation, assume first that S_t^0 is fixed at some value S^0. This level of real spending can be financed by different mixes of inflationary finance on the one hand, and taxation, debt issue, etc., on the other, where the amount to be collected by money issue is expressed as $k_t S^0$. In the usual diagram plotting the demand for real balances as a function of the inflation rate, $k_t S^0$ is measured in steady states by the area of the rectangle defined by μ—the inflation rate—and M_t^0/e_t^0.

Consider now an increment in μ. Real balances decline according to the money demand function, and the revenue from inflation, $k_t S^0$, increases as long as μ is below the rate that corresponds to a unitary demand elasticity for real balances. Because real spending is constant, k_t increases, and, hence, the fraction of S^0 financed by other means declines.

This shift from taxation to money issue can be viewed as the policy variable that brings about higher inflation rates. Classic works on the German hyperinflation, like those by Graham (1930) and Bresciani-Turroni (1937), describe an opposite direction of effect. Namely, the rate of depreciation of the currency had a negative effect on the real yield from taxation due to the interval of time existing between the occurrence of taxable transactions and the actual payment of the taxes.[8] The present discussion relies on the correlation between the fraction of expenditure financed by money issue and the inflation rate rather than on a specific mechanism relating these two variables. This positive correlation implies that k_t and $1/(M_t^0/e_t^0)$ move in the same direction. However, this coincidental movement does not hold for all μ. When μ reaches the rate that maximizes the revenue from inflation, k_t also reaches its highest level, and when it rises above that rate, k_t must decline. In other words, the correlation between k_t and $1/(M_t^0/e_t^0)$ turns negative in that range.

This decline in k_t implies that the revenue from other sources must go up. If tax collection and debt issue cannot be increased (e.g., due to the negative effect of inflation mentioned above), spending must be partially financed by extraordinary means, such as sales of gold from

[8] See, e.g., Bresciani-Turroni 1937, p. 66; and Graham 1930, p. 44.

the Central Bank's stock. In fact, the balance sheet of the German Central Bank shows that the stock of gold begins to decline significantly in April 1923, after being fairly stable since 1920.[9]

Given this behavior of k_t when S_t^0 is constant, equation (18) can be approximated by the semilogarithmic form:

$$\frac{M_t^0 - M_{t-1}^0}{M_t^0} = \text{constant} + b' \log\left(\frac{1}{M_t^0/e_t^0}\right) + u', \qquad (19)$$

where b' is a positive coefficient, u' is a random term of zero mean, and the constant term is affected by the level of S^0. In this specification, the implicit fraction k_t increases along with $1/(M_t^0/e_t^0)$ at lower and middle ranges of this variable but eventually declines when real balances fall below a certain value.

Equation (19) acquires more empirical content if real spending is allowed to vary. During the period under study, S_t^0 has a declining trend. However, holding constant the real balances term, which has a strong correlation with time, movements in S_t^0 can be interpreted as temporary deviations from a "normal" trend. These fluctuations are assumed also to be correlated with the fraction k_t. The assumption here is that given relatively high costs associated with temporary shifts in tax collection and debt issue, transitory movements in spending would be financed primarily by adjustments in money issue. A positive correlation between S_t^0 and k_t would then result. However, a sufficiently high value of S_t^0 could be presumed to require extraordinary finance of the sort previously mentioned, so that k_t might eventually decline.

Incorporating an approximation of this effect into equation (19) results in the following generalized expression:

$$\frac{M_t^0 - M_{t-1}^0}{M_t^0} = a' + b' \log\left(\frac{1}{M_t^0/e_t^0}\right) + c' \log S_t^0 + u',$$

which can be rewritten as

$$\frac{M_t^0 - M_{t-1}^0}{M_t^0} = a' - b'[M_{t-1} + (M_t - M_{t-1}) - e_t] + c'S_t + u', \qquad (20)$$

where variables without a superscript are again in logarithmic terms.

In order to proceed with the formulation of the prior expectation, it is convenient to replace the logarithmic growth rate $(M_t - M_{t-1})$ on the right-hand side by the growth rate measured by $(M_t^0 - M_{t-1}^0)/M_t^0$. While this rate is always lower than $M_t - M_{t-1}$, the gap widening the higher the growth rates, this effect can hopefully be captured approximately by the coefficients in the estimated equation. Then,

[9] See Statistisches Reichsamt 1925, p. 53.

equation (20) can be solved for $(M_t^0 - M_{t-1}^0)/M_t^0$ to yield

$$\frac{M_t^0 - M_{t-1}^0}{M_t^0} = \frac{a}{1+b} - \frac{b}{1+b}(M_{t-1} - e_t) + \frac{c}{1+b}S_t + \frac{1}{1+b}u_t. \quad (21)$$

The prior conditional expectation is defined accordingly as

$$g_t \equiv \left(\frac{\hat{a}}{1+b}\right) - \left(\frac{\hat{b}}{1+b}\right)(M_{t-1} - e_t) + \left(\frac{\hat{c}}{1+b}\right)S_t. \quad (22)$$

The unperceived part of money growth $\tilde{m}_t$[10] is then computed by the difference between actual growth and g_t, namely,

$$\tilde{m}_t = (M_t^0 - M_{t-1}^0)/M_t^0 - g_t.$$

The coefficients in equation (22) are those which result from regressing $(M_t^0 - M_{t-1}^0)/M_t^0$ on S_t and $(M_{t-1} - e_t)$. However, the exchange rate is not in general an exogenous variable in a money growth equation. A correlation between e_t and the error term u_t will exist via some unspecified condition for equilibrium in the foreign exchange market. Therefore, the coefficients in (22) do not correspond exactly to those in equation (21). This property is not a drawback. On the contrary, the bias in the estimated coefficients (relative to those in [21]) reflects the part of u_t that can be estimated from e_t. It therefore should be taken into account in calculating g_t.[11]

Turn now to the estimation of equation (21). There is a problem in matching the available data on money with those on prices for the German hyperinflation. Unlike the price series, which consist of monthly averages, the available data on the money stock until January 1923 are end-of-month figures.[12] From January 1923 onward, four quotations per month are available. Thus, a proxy is constructed for the monthly average money stock. For the period January–July 1923 it contains averages of the beginning-of-month, end-of-month, and the three intermediate quotations available. Until December 1922, the monthly averages are approximated by linear interpolation of the end-of-month figures. The constructed series are shown in table 2.

A further consideration arises. The estimation of equation (21) from monthly average data on the money stock, rather than end-of-

[10] The unperceived growth $\tilde{m}_t$ does not correspond exactly to the error term u_t. On this point, see below.

[11] E.g., in the general linear model $y = X\beta + u$, where the variables in X are correlated with u, the estimated vector of coefficients is $\hat{\beta} = (X'X)^{-1}X'y = (X'X)^{-1}X'(X\beta + u)$, or $\hat{\beta} = \beta + (X'X)^{-1}X'u$, where $(X'X)^{-1}X'u$ is the regression coefficient of u on X. The prediction of y_t given the values of the vector x_t is, accordingly, $y_t = x_t'[\beta + (X'X)^{-1}X'u]$; i.e., it is composed of the systematic part $x_t'\beta$, plus the conditional expectation of u_t given x_t.

[12] Another problem stems from the form of the prior expectation, which, as it stands, requires the use of end-of-month money stocks. See the discussion below.

TABLE 2

VALUES (in Logarithms) OF REAL EXPENDITURE, EXCHANGE RATE, AND MONEY STOCK

	S_t	e_t	M_t
Oct. 1920	5.85	2.79	11.24
Nov.	6.44	2.91	11.25
Dec.	6.58	2.86	11.28
Jan. 1921	6.42	2.74	11.29
Feb.	6.80	2.68	11.28
Mar.	6.66	2.70	11.29
Apr.	6.65	2.71	11.30
May	6.53	2.69	11.30
June	6.76	2.80	11.33
July	6.43	2.91	11.36
Aug.	6.76	3.00	11.38
Sept.	6.23	3.22	11.42
Oct.	5.95	3.58	11.48
Nov.	5.54	4.14	11.55
Dec.	6.44	3.82	11.66
Jan. 1922	5.96	3.82	11.72
Feb.	5.84	3.90	11.75
Mar.	5.85	4.22	11.81
Apr.	5.77	4.24	11.88
May	5.93	4.24	11.96
June	5.75	4.33	12.05
July	5.34	4.77	12.16
Aug.	5.32	5.60	12.33
Sept.	6.08	5.86	12.58
Oct.	5.60	6.63	12.92
Nov.	5.30	7.44	13.35
Dec.	6.08	7.50	13.85
Jan. 1923	5.32	8.36	14.26
Feb.	5.63	8.80	14.82
Mar.	6.55	8.53	15.31
Apr.	6.15	8.67	15.60
May	5.65	9.34	15.83
June	6.21	10.17	16.29
July	6.16	11.34	17.14

NOTE.—S_t = log of the monthly government expenditure in millions of gold marks. Source: Bresciani-Turroni (1937), pp. 436–37. e_t = log of the monthly average exchange rate of the gold marks in paper marks. Source: Bresciani-Turroni (1937), p. 441. M_t = log of the monthly average money stock in millions of paper marks. Source: based on fixed days quotations from *Sonderhefte zur Wirtschaft und Statistic* (1925), pp. 45–47.

month figures, means that both the current month's spending and that of the previous month should be considered. Spending financed by money issue during the previous month increases one to one the current monthly average stock but has, in general, a weaker effect on the prior month's average stock.[13] To account for this effect, lagged

[13] This effect can be seen by considering first a case where each month's spending is spread evenly over the month, and, say, it is financed only by the issue of new money. In this case the money stock grows linearly at, in general, different rates within each monthly period. Then, if $M^0_{t-2,\text{end}}$ denotes the money stock at the end of month $t-2$,

spending is incorporated into the framework of the semilogarithmic function in equation (21). First, define the variables S_t' and e_t' by $S_t' \equiv \log [\xi S_t^0 + (1 - \xi)S_{t-1}^0]$ and $e_t' \equiv \log [\xi S_t^0 + (1 - \xi)S_{t-1}^0]$. After substituting S_t' and e_t' for S_t and e_t in (21), ξ is estimated, simultaneously with the other coefficients in the equation, using a nonlinear maximum likelihood procedure under normally distributed errors.[14]

The estimated nonlinear equation is

$$\frac{M_t^0 - M_{t-1}^0}{M_t^0} = \underset{(.071)}{.660} - \underset{(.007)}{.166}\,(M_{t-1} - e_t') + \underset{(.014)}{.115}S_t' \tag{23}$$

$$(\hat{\xi} = \underset{(.11)}{.75}, R^2 = .98, \text{D-W} = 1.63, \sigma = .026, 33 \text{ observations}),$$

where the numbers in parentheses are the standard errors of the coefficients. With respect to the number of observations, the starting month was taken as November 1920 in order to test lagged effects of monetary shocks on price dispersions. According to the argument of note 13 above, the value .75 for ξ suggests a pattern of spending that is biased toward the beginning of the month. In order to proceed with the analysis on the more familiar ground of linear equations estimation, ξ is assumed henceforth to equal .75. Given this value, the standard errors of the other coefficients, linearly estimated, are not materially different from those obtained above. In order of their appearance in (23), they are .068, .005, and .014. The standard error of the regression is now 0.0256. The other regression statistics and the coefficients remain obviously the same.

The pattern of the residuals from equation (23), which are reported in table 4, suggests that their variance increased during the sample period as inflation progressed. A method similar to that proposed by Glejser (1969) is adopted to correct for the apparent heteroscedasticity by assuming a specific model for the variance of the error term. In this procedure the variance is postulated to be determined by a set of variables $\{z_i\}$ in the linear form:

$$\sigma_{mt}^2 = \sum_i \omega_i z_{it}. \tag{24}$$

---

the monthly average for $t - 1$ equals $M_{t-2,\text{end}} + \frac{1}{2} S_{t-1}^0 e_{t-1}^0$, and that corresponding to month t equals $M_{t-2,\text{end}}^0 + S_{t-1}^0 e_{t-1}^0 + \frac{1}{2} S_t^0 e_t^0$. Thus, the increase in the monthly average from $t - 1$ to t equals $\frac{1}{2} S_{t-1}^0 e_{t-1}^0 + \frac{1}{2} S_t^0 e_t^0$. Namely, spending evenly spread over each month would imply equal weights for current and lagged spending in the money growth equation. Alternatively, if spending is concentrated at the beginning of the month, the relative weight of lagged expenditure would be lower. At the extreme, e.g., if all spending is made only on the first day of each month, both $t - 1$ and t monthly averages increase equally by the amount of the $t - 1$ expenditure. In this case lagged spending does not belong in the money growth equation.

[14] This procedure is from the TSP Regression Package.

If the values of the true money shocks $\tilde{m}_t^*$ were available, one could use them as follows. Since the expectation of $\tilde{m}_t^{*2}$ is σ_{mt}^2, it follows that

$$\tilde{m}_t^{*2} = \sigma_{mt}^2 + v_t,$$

where v_t is of zero mean. Combining the last two equations yields

$$\tilde{m}_t^{*2} = \sum_i \omega_i z_{it} + v_t. \tag{25}$$

Estimates of the coefficients in equation (24) could be obtained by regressing $\tilde{m}_t^{*2}$ on the z_i variables. The heteroscedasticity problem is also present here, but it will be ignored in what follows.

The values of $\tilde{m}_t^*$ are, however, unknown; only the estimated residuals $\tilde{m}_t$ are available from the OLS money growth equation. Since $\tilde{m}_t^2$ converges to $\tilde{m}_t^{*2}$ asymptotically, the variance estimated using $\tilde{m}_t^2$ values obtained from small samples will be biased. This bias is neglected, as it is probably of relatively small importance.

In order to proceed with the implementation of this procedure the set of the z_i variables must be specified. The presumption is that the same variables used to explain the growth rates are also correlated with the variances. Thus, S_t', e_t', and M_{t-1} are candidates. The lagged squared residual $\tilde{m}_{t-1}^2$ is also included as an explanatory variable. It presumably captures the effect of serially correlated omitted variables and perhaps a direct correlation between σ_{mt-1}^2 and the current variance. The estimated variance equation is

$$\begin{aligned} \tilde{m}_t^2 = &- .007 - .00038S_t' - .00026e_t' + .00090M_{t-1} - .281\tilde{m}_{t-1}^2 \\ &(.003)\ (.00047)\quad (.00031)\quad (.00047)\qquad (.130) \end{aligned} \tag{26}$$

$$(R^2 = .57,\ \text{D-W} = 2.5,\ \sigma = .001,\ 33 \text{ observations}).^{15}$$

Nothing in this procedure for estimating the series of money variances guarantees that all the fitted values from equation (26) would be positive. Indeed, two of the fitted values have a negative sign. In order to use the estimated series as a measure of variances, these two negative values are replaced with the smallest positive value in the series. The series of the square roots of these estimates are reported in column 4 of table 4.

Using this series as weights for the corresponding observations, equation (23) is reestimated with the following results:

$$\begin{aligned} (M_t^0 - M_{t-1}^0)/M_t^0 = &.700 - .154(M_{t-1} - e_t') + .094S_t' \\ &(.047)\ (.006)\qquad\qquad (.012) \end{aligned} \tag{27}$$

$$(R^2 = .95,\ \text{D-W} = 1.8,\ \sigma = 1.01,\ 33 \text{ observations}).$$

[15] In order to estimate the 33 variances needed to reestimate eq. (34), the $\tilde{m}_t$ series were obtained from running the money growth equation after adding the additional observation of October 1920.

Observe that the coefficient of S_t' here is somewhat lower than in the OLS equation and that the coefficient of $(M_{t-1} - e_t')$ is somewhat higher. The general form of the equation is, however, robust to this transformation of the data.

The next step is to test the stability of the coefficients in equation (27) across two subperiods. Stability of the coefficients has particular relevance here. If the equation is approximately stable, it seems easier to assume that it was known from the beginning of the period and that perceptions about money growth were formed using the same equation during the entire sample.[16] The period is divided into an approximately nonaccelerating money supply period until May 1922 and an accelerating phase beginning in June 1922. An F-test applied to these two subperiods yields the statistic $F^3_{27} = 1.7$, with a corresponding 5 percent critical value of 3.0. Therefore, the hypothesis of stable coefficients across these two subperiods cannot be rejected at the 5 percent significance level.

A regression in which the coefficients of M_{t-1} and e_t' are unconstrained produces coefficients of similar magnitude for the two variables. The F-test for the linear constraint of equal coefficients yielded the statistic $F^1_{29} = 2.3$, where the 5 percent critical value is 4.2.

IV. Empirical Test of the Dispersion Equation

The tests of the price dispersion model in equation (15) are performed using the dispersion series computed in Section II and the unperceived monetary shocks as measured by the residuals in equation (27).

For convenience equation (15) is rewritten here

$$\gamma_t^2 = \left\{2(1-\theta)^2\sigma_\lambda^2 + 2[\theta + \lambda(1-\theta)]^2\right\}\sigma_\epsilon^2 + (1-\theta)^2\sigma_\lambda^2(\tilde{m}_t - \tilde{m}_{t-1})^2. \tag{15}$$

Significantly, this equation has a simple linear form—and can therefore be tested by an ordinary least-squares procedure—only under constant money and relative shocks variances. However, the analysis of money growth in the previous section suggested that σ_m^2 increased during the hyperinflation. If this is indeed the case, it would not be appropriate to test the model with a specification that relates price dispersion to money shocks with a constant coefficient.

Two different procedures are adopted here to deal with the possibility of a changing money variance. The first uses a linear approximation in which the variance of money—as measured by the $\hat{\sigma}_m^2$ series

[16] Estimation of a money growth equation in a similar context, using the entire sample (for U.S., 1941–73), was discussed and performed in Barro (1977).

from Section III—is kept constant by including it additively in the equation. As discussed in Section I, σ_m^2 has different and opposite effects on γ_t^2, and, therefore, on an a priori basis the coefficient of $\hat{\sigma}_m^2$ could take either sign. The other attempt is to estimate (15) as a nonlinear equation. The results of this procedure, reported later in the section, are quite poor.

The estimated equation in which $\hat{\sigma}_m^2$ is added linearly is:

$$\gamma_t^2 = \underset{(.010)}{.033} + \underset{(3.2)}{17.4}(\bar{m}_t - \bar{m}_{t-1})^2 - \underset{(8.6)}{15.8}\hat{\sigma}_{mt}^2 \tag{28}$$

$$(R^2 = .59,\ \text{D-W} = 1.2,\ \sigma = .022,\ 30 \text{ observations}).$$

The monetary shocks appear to have considerable explanatory power for price dispersion. The coefficient of $\hat{\sigma}_{mt}^2$ is negative and, therefore, suggests a dominant Lucas-type effect of the money variance on price dispersion. That is, the degree of dispersion associated with given shocks diminishes the higher their variance. The explanatory power of $\hat{\sigma}_{mt}^2$, however, is fairly low; its coefficient is significantly different from zero at the 5 percent level but fails to be so at the 2.5 percent level.

Theoretically, the 1-month lagged money variance $\hat{\sigma}_{mt-1}^2$ belongs also in the equation. However, when included, its coefficient is insignificant with a t-statistic of .8.

The Durbin-Watson statistic indicates autocorrelated residuals, which may be caused by omitted real variables (like changes in the pattern of government spending,[17] in income distribution, etc.) that are serially correlated or by the fact that the $\bar{m}_t$ variable includes an estimation error. In order to check whether the degree of significance of the estimated coefficients in (28) is affected by this autocorrelation, the equation is reestimated using the Cochrane-Orcutt technique. The results are quite similar to those obtained before:

$$\gamma_t^2 = \underset{(.008)}{.038} + \underset{(2.9)}{16.8}(\bar{m}_t - \bar{m}_{t-1})^2 - \underset{(9.8)}{18.5}\hat{\sigma}_{mt}^2 \tag{29}$$

$$(R^2 = .67,\ \text{D-W} = 2.2,\ \sigma = .020,\ 29 \text{ observations},\ \hat{\rho} = \underset{(.17)}{.36}).$$

Including $\hat{\sigma}_{mt-1}^2$ in this regression yielded a t-statistic of only 1.0 for its coefficient. The possibility of lagged effects of monetary shocks on price dispersion was explored by including the variable $(\bar{m}_{t-1} - \bar{m}_{t-2})^2$ in the equation, but the estimated coefficient was found statistically insignificant. The t-ratio was 0.3 in the OLS regression and 1.4 using the Cochrane-Orcutt technique.

[17] The magnitude of changes in the amount of real government spending, however, does not have any significant explanatory power.

The next step is to see whether the dispersion equation is stable over the entire period. In fact, the inclusion of the $\hat{\sigma}^2_{mt}$ variable is an attempt to control one source of instability in the coefficients. In order to carry out this test the sample is divided first after May 1922, estimating the equation separately for the two subperiods. This partition of the sample is the same one adopted previously to test the stability of the money growth equation. Then, the exercise is repeated, dividing the sample at the end of 1922. The aim of this partition is to see how the model performs after removing from the sample the 7 months of 1923, which had a much more unstable monetary growth.

The results of these regressions, reported in table 3, can be summarized as follows. When the sample is divided in May 1922 (first column), the results for the first subperiod are fairly weak. Both coefficients are insignificant in the OLS equation. Using the Cochrane-Orcutt technique the coefficient of $(\tilde{m}_t - \tilde{m}_{t-1})^2$ turns out significant at the 5 percent level, although that corresponding to $\hat{\sigma}^2_{mt}$ is still insignificant. During the second subperiod—from June 1922 to July 1923—the statistical performance of the equation is much stronger.

The second column reports the equations estimated for the periods through and after December 1922. Observe that the removal of the 1923 portion of the sample worsens the performance of the model as judged by the size of the t-ratios. However, the coefficient of the monetary shocks still remains quite significant.

Formal F-tests fail to reject the hypothesis of stable coefficients across the mentioned subperiods. When the sample is divided in May 1922, the resulting statistic is $F^3_{24} = 1.8$, while the 5 percent critical value is 3.0. Partitioning the sample at the end of 1922 yields the statistic of only 0.8.

The other approach adopted to test the dispersion equation was to treat it as a nonlinear relationship. When σ^2_m is changing over time, equation (15) generalizes to

$$\begin{aligned}\gamma^2_t = {} & \sigma^2_\epsilon \sigma^2_\lambda \left[\left(\frac{\sigma^2_\epsilon}{\lambda\sigma^2_{mt} + \sigma^2_\epsilon}\right)^2 + \left(\frac{\sigma^2_\epsilon}{\lambda\sigma^2_{mt-1} + \sigma^2_\epsilon}\right)^2\right] \\ & + \sigma^2_\epsilon\left[\left(\frac{\lambda\sigma^2_{mt} + \lambda\sigma^2_\epsilon}{\lambda\sigma^2_{mt} + \sigma^2_\epsilon}\right) + \left(\frac{\lambda\sigma^2_{mt-1} + \lambda\sigma^2_\epsilon}{\lambda\sigma^2_{mt-1} + \sigma^2_\epsilon}\right)\right] \\ & + \sigma^2_\lambda\left(\frac{\sigma^2_\epsilon}{\lambda\sigma^2_{mt} + \sigma^2_\epsilon}\tilde{m}_t - \frac{\sigma^2_\epsilon}{\lambda\sigma^2_{mt-1} + \sigma^2_\epsilon}\tilde{m}_{t-1}\right)^2.\end{aligned} \tag{30}$$

As discussed in Section I, the third term in this equation reflects the negative effect of σ^2_m on the impact of monetary shocks. The second term corresponds to Barro's (1976) relative price variance expression

TABLE 3

ESTIMATED PRICE CHANGE DISPERSION EQUATIONS

	Feb 1921–May 1922	Feb. 1921–Dec 1922	Feb. 1921–July 1923
OLS	$.030 + 18.3(\dot{m}_t - \bar{\dot{m}}_{t-1})^2 - 28.2\hat{\sigma}^2_{mt}$ (.012) (15.8) (32.2) (R^2 = .22, D-W = 1.2, σ = .019)	$.024 + 34.4(\dot{m}_t - \bar{\dot{m}}_{t-1})^2 - 7.3\hat{\sigma}^2_{mt}$ (.009) (12.1) (19.0) (R^2 = .29, D-W = 1.1, σ = .022)	$.033 + 17.4(\dot{m}_t - \bar{\dot{m}}_{t-1})^2 - 15.8\hat{\sigma}^2_{mt}$ (.010) (3.2) (8.6) (R^2 = .59, D-W = 1.2, σ = .022)
Cochrane-Orcutt	$.035 + 27.7(\dot{m}_t - \bar{\dot{m}}_{t-1})^2 - 43.1\hat{\sigma}^2_{mt}$ (.012) (13.5) (32.0) (R^2 = .47, D-W = 2.5, σ = .016)	$.041 + 30.7(\dot{m}_t - \bar{\dot{m}}_{t-1})^2 - 35.1\hat{\sigma}^2_{mt}$ (.012) (9.0) (21.2) (R^2 = .53, D-W = 2.2, σ = .018)	$.038 + 16.8(\dot{m}_t - \bar{\dot{m}}_{t-1})^2 - 18.5\hat{\sigma}^2_{mt}$ (.008) (2.9) (9.8) (R^2 = .67, D-W = 2.2, σ = .020)
	June 1922–July 1923	Jan. 1923–July 1923	
OLS	$.054 + 17.5(\dot{m}_t - \bar{\dot{m}}_{t-1})^2 - 28.1\hat{\sigma}^2_{mt}$ (.012) (3.6) (11.6) (R^2 = .70, D-W = 1.6, σ = .022)	$.036 + 16.6(\dot{m}_t - \bar{\dot{m}}_{t-1})^2 - 16.4\hat{\sigma}^2_{mt}$ (.026) (3.2) (15.0) (R^2 = .87 D-W = 3.1, σ = .020)	

that is affected positively by σ_m^2 when $0 < \lambda < 1$. Recall that λ is defined as the average of $1/[\alpha^s(z) + \alpha^d]$ across markets, and $\alpha^s(z) + \alpha^d$ is the excess demand elasticity of commodity z. Using again the $\hat{\sigma}_m^2$ series, the parameters of equation (30)—λ, σ_ϵ^2, and σ_λ^2—are estimated using a nonlinear least-squares procedure with fairly weak results. The additional structure given to the equation seems to be rejected by the data, as judged by a higher sum of squared errors than in the linear equation.

Observe that a value of zero for λ reduces the equation to the linear form of equation (15), in which the money variance is constant. The estimated value for this parameter was 0.056 with a standard error of 0.065.[18] The interpretation of this result is not that the average $1/[\alpha^s(z) + \alpha^d]$ is likely to be close to zero. Instead, it suggests that the detailed specification of equation (30) is too stringent. For example, if the variance of the relative excess demand shifts, σ_ϵ^2 also changes over time: this is more of a problem in this approach, since σ_ϵ^2 itself is being estimated as a constant.

V. Actual Money Growth and Inflation in the Dispersion Equation

In this section additional variables that were mentioned in the literature as being related to price dispersion are tried in the equation. There is no rigorous theoretical justification for their inclusion. Thus, only loose verbal explanations are given. Also, the variables related to inflation are clearly not exogenous, and, therefore, any observed correlation cannot imply causality.

Actual money growth and price dispersion.—A variable that can be considered exogenous and, if one assumption of Section I is violated, in principle also relevant for price dispersion is the actual money growth. Changes in the money stock can affect the dispersion of prices (even when perceived) if the new money is spread unevenly across the economy, thereby affecting relative demand in different sectors. This type of effect was discussed by Cairnes (1873) with respect to gold discoveries. In the framework of the present model this type of effect could be represented by changes in the relative excess demand variance σ_ϵ^2. Since here the focus is on price change dispersion, the corresponding variable in this context is the change in the growth rate or the degree of acceleration/deceleration in the money stock.

In order to test this sort of effect, and also to see whether ($\bar{m}_t$ –

[18] $\hat{\sigma}_\lambda$ was 17.7 with a standard error of 8.4 and $\hat{\sigma}_\epsilon^2$ was 0.0008, with a standard error of 0.0004.

$\bar{m}_{t-1})^2$ is only a proxy for changes in actual money growth, price dispersion was regressed on $[(M_t^0 - M_{t-1}^0)/M_t^0 - (M_{t-1}^0 - M_{t-2}^0)/M_{t-1}^0]^2$ with the following results:

$$\gamma_t^2 = \underset{(.006)}{.033} + \underset{(.6)}{2.1}\left(\frac{M_t^0 - M_{t-1}^0}{M_t^0} - \frac{M_{t-1}^0 - M_{t-2}^0}{M_{t-1}^0}\right)^2$$

($R^2 = .33$, D-W = 1.5, $\sigma = .028$, 30 observations).

The monetary acceleration/deceleration variable has a statistically significant correlation with price dispersion. Remarkably, however, its explanatory power vanishes when $(\bar{m}_t - \bar{m}_{t-1})^2$ and $\hat{\sigma}_{mt}^2$ are also included in the regression. The equation including the three variables is

$$\gamma_t^2 = .034 + \underset{(4.0)}{16.6}(\bar{m}_t - \bar{m}_{t-1})^2 - \underset{(9.0)}{16.5}\hat{\sigma}_{mt}^2 +$$

$$\underset{(.77)}{.25}\left(\frac{M_t^0 - M_{t-1}^0}{M_t^0} - \frac{M_{t-1}^0 - M_{t-2}^0}{M_{t-1}^0}\right)$$

($R^2 = .59$, D-W = 1.3, $\sigma = .022$, 30 observations).

This result denies the existence of any effect of relative demands following the introduction of new money during this period. It supports the hypothesis that money affects relative prices only if it is currently unperceived.

Inflation and price dispersion.—In his analysis of price behavior during the hyperinflation, Graham describes a positive correlation between the acceleration of price level and the dispersion of prices. Mills's findings in his study of U.S. prices suggest that dispersion is positively correlated with both acceleration and deceleration of the price level.

If the acceleration/deceleration in the price level is related to an unperceived monetary expansion or contraction, the theory tested here predicts that the correlation mentioned above should be captured by a variable measuring unperceived money growth. To test whether this is the case here, the variable $(\mu_t - \mu_{t-1})^2$ is also included in the equation—where μ_t is the inflation rate from month $t - 1$ to month t computed from the wholesale price index (see table 4). The results suggest that there is a separate correlation between $(\mu_t - \mu_{t-1})^2$ and price dispersion. The equation estimated by OLS is

$$\gamma_t^2 = \underset{(.005)}{.032} + \underset{(3.0)}{18.6}(\bar{m}_t - \bar{m}_{t-1})^2 - \underset{(8.7)}{24.2}\hat{\sigma}_{mt}^2 + \underset{(.031)}{.074}(\mu_t - \mu_{t-1})^2$$

($R^2 = .67$, D-W = .9, $\sigma = .020$, 30 observations).

Given the low D-W statistic, the equation was reestimated by

TABLE 4

VALUES OF MONEY GROWTH AND INFLATION

	$(M_t^0 - M_{t-1}^0)/M_t^0$ (1)	$\widehat{(M_t^0 - M_{t-1}^0)/M_t^0}$ (2)	Col. 1 − Col. 2 (3)	$\hat{\sigma}_m$ (4)	g_t (5)	$\hat{m}_t$ (6)	μ_t (7)
Nov. 1920	.010	003	.008	.008*	.006	.004	.029
Dec.	.026	.025	001	.011	.024	.002	−.047
Jan. 1921	.008	− 006	.014	.014	−.003	.011	.001
Feb.	−.011	.010	−.021	010	.008	−.019	−.045
Mar.	.010	.009	.001	.008*	.008	.002	−.028
Apr.	.007	.005	.002	.012	.004	.003	−.009
May	.008	−.009	.017	.014	−.007	.015	−.014
June	022	019	.003	.009	.017	.005	.043
July	.029	.012	.017	.014	.012	.016	.044
Aug.	.022	.042	−.020	.009	.038	−.016	.294
Sept.	.045	.036	009	.008	.036	.009	.075
Oct.	.057	.041	.016	.017	.046	.011	.174
Nov.	.069	.076	−.007	.018	.083	−.014	.328
Dec	.100	.117	−.017	.013	.113	−.013	.021

Jan. 1922	.061	065	−.003	017	.067	−.006	.050
Feb.	.024	.037	−.013	.023	.044	−.021	.113
Mar.	060	.074	−.014	.020	.079	−.019	.281
Apr.	.075	.072	.003	.021	.078	−.003	.158
May	.070	.071	−.001	.024	.076	−.006	.015
June	.088	.059	028	026	.066	.022	.085
July	.106	.066	.040	.025	.077	.029	.358
Aug.	.158	.161	−.002	.020	.166	−.008	.646
Sept.	.221	.260	−.038	.023	251	−.029	.402
Oct.	.285	.310	−.025	.016	.300	−.015	.679
Nov.	.349	.347	.002	.027	339	.010	.712
Dec.	.393	.372	.021	.032	.355	.038	.245
Jan. 1923	.339	.367	−.027	.036	.355	−.016	.636
Feb.	.431	.380	050	.038	.367	.063	.696
Mar.	.385	.364	.021	.034	.342	043	−.133
Apr.	.252	.275	−.024	.047	.261	−.009	.064
May	.205	268	−.064	.050	.260	−.056	.450
June	371	.399	−.028	.037	.378	−.007	.864
July	.572	.518	.054	.049	.488	.084	1.350

NOTE.—$\widehat{(M_t^s - M_{t-1}^s)/M_t^s}$ is the estimated value of money growth from eq. (23). The data used in the estimation are shown in table 2. $\hat{\sigma}_m$ = square root of the estimated value from eq. (26). The entries with an asterisk are those with a negative fitted value that were replaced by the smallest positive value in the series. g_t = estimated value from the weighted least-squares regression (eq. [27]), where the series $\hat{\sigma}_m$ are used as weights. $\hat{m}_t = (M_t^s - M_{t-1}^s)/M_t^s - g_t$. μ_t = first difference of the logs of the Wholesale Price Index. Data obtained from *Sonderhefte zur Wirtschaft und Statistic* (1925).

Cochrane-Orcutt:

$$\gamma_t^2 = \underset{(.008)}{.033} + \underset{(2.6)}{19.5}(\bar{m}_t - \bar{m}_{t-1})^2 - \underset{(9.0)}{25.3}\hat{\sigma}_{mt}^2 + \underset{(.022)}{.077}(\mu_t - \mu_{t-1})^2$$

$$(R^2 = .77,\ \text{D-W} = 2.00,\ \sigma = .017,\ 29 \text{ observations},\ \hat{\rho} = \underset{(.16)}{.49}).$$

Not only does $(\mu_t - \mu_{t-1})^2$ have a statistically significant correlation with price dispersion, but its inclusion in the equation also sharpens the performance of the original two variables. Possible explanations of this correlation could be related, for example, to income redistribution following from unanticipated inflation or substitution between money and certain commodities as stores of value, when their relative costs change.

Sheshinski and Weiss (1977) consider a model of a monopolistic firm in which costs involved in changing the price of the commodity produced generate discrete periodic price adjustments whose magnitude increases with the inflation rate. They suggest that, if the timing of these adjustments is independent across firms, higher inflation rates imply larger dispersion of individual price changes. Hence, this type of argument seems to imply a relationship between price change dispersion and μ_t^2.[19]

A role for μ_t^2 may be rationalized also on different grounds. Graham mentions the effect of the interaction between government regulations and inflation on the dispersion of prices. He refers to price controls, such as rent restriction legislation, which generate dispersion among the differently affected prices as the general price level increases. According to this argument, the dispersion of prices will be positively correlated with the price level. Hence, the dispersion of price changes will increase with the magnitude of changes in the price level—namely with μ_t^2.

Adding μ_t^2 to the equation yields the following results:

[19] However, the empirical implications of Sheshinski and Weiss's analysis for price dispersion do not seem clear to me. An ambiguity arises because of the probably positive effect of inflation on the frequency of price changes that they derived. If the length of the observation period is kept constant, a higher frequency of price change may diminish the measured dispersion of price changes. The possibility of a negative effect of inflation on price dispersion in this framework can be seen in the following example. Assume that the optimal frequency of price adjustments for all firms is 2 months and that part of the firms adjust their prices during odd-numbered months and the rest during even-numbered months. Using monthly data, dispersion of price changes will depend on the magnitude of price changes corresponding to the group of firms currently adjusting prices. Now assume that inflation increases; as a consequence the magnitude of price adjustments goes up, and also the optimal frequency is increased—say to one per month. Since now all the firms adjust prices during the same month, the dispersion of price changes collapses to zero, in spite of the larger individual price changes.

$$\gamma_t^2 = \underset{(.005)}{.031} + \underset{(4.1)}{12.2}(\bar{m}_t - \bar{m}_{t-1})^2 - \underset{(8.2)}{21.3}\hat{\sigma}_m^2 + \underset{(.030)}{.062}(\mu_t - \mu_{t-1})^2 + \underset{(.017)}{.036}\mu_t^2$$

$$(R^2 = .67,\ \text{D-W} = 1.3,\ \sigma = .019,\ 30 \text{ observations}).$$

The coefficients of μ_t^2 and $(\mu_t - \mu_{t-1})^2$ are both significant at the 2.5 percent level. However, when the equation is reestimated by Cochrane-Orcutt, the effect of μ_t^2 becomes insignificant (t-ratio of 0.8), while the performance of the rest of the variables improves. If the monetary acceleration/deceleration variable is also included, its coefficient turns out negative but insignificant at the 5 percent level (t-ratio of 1.62 with OLS and 1.55 with Cochrane-Orcutt), without materially affecting the other coefficients and standard errors.

VI. Summary and Conclusions

This paper tested a model of price dispersion using data on the German hyperinflation after World War I. The theoretical framework, outlined in Section I, predicted that money growth causes price dispersion only if it is currently unperceived.

In order to perform this test, the first task was to construct a measure of price dispersion based on a broad enough range of commodities to approximate the varying degree of relative price variability during that period. This calculation was accomplished using wholesale price data on a sample of over 60 commodities, including foodstuffs, textiles, metals, and fuels.

The next step was the delicate one of measuring the perceived part of money growth. This estimation implied the postulation of an information set available at each point in time and also a function relating the variables in this set to the creation of money. The function used to isolate the perceived part of money growth was specified based on considerations related to the government demand for revenue to finance expenditure.

A measure of the variance of the unperceived money growth was also estimated. This series was used for dealing with the heteroscedasticity problem in the money growth equation and also as a theoretically relevant factor affecting price dispersion.

The testing of the dispersion equation showed a statistically significant correlation between unperceived money—as measured here—and price dispersion. Thus, the model is supported to some extent by the empirical results. The variance of the money shocks turned out to have a negative coefficient which was only marginally significant. Although the money variance has different and opposite theoretical effects on price dispersion, this negative correlation is consistent with a dominant effect of diminishing the misperceptions of

monetary disturbances as relative shifts—and thus reducing the dispersion associated with those disturbances.

In Section V actual money growth was also considered as a price dispersion factor, justified as affecting relative demands. However, the results indicated no additional effect for the rate of money creation. The magnitude of changes in the inflation rate does appear to have a separate correlation with price change dispersion. The present model, however, does not provide an explanation for this association.

References

Barro, Robert J. "Rational Expectations and the Role of Monetary Policy." *J. Monetary Econ.* 2 (January 1976): 1–32.

———. "Unanticipated Money Growth and Unemployment in the United States." *A.E.R.* 67 (March 1977): 101–15.

Bresciani-Turroni, Costantino. *The Economics of Inflation.* London: Allen & Unwin, 1937.

Cairnes, John E. "The Course of Depreciation." In *Essays in Political Economy: Theoretical and Applied.* London: Macmillan, 1873.

Glejser, H. "A New Test for Heteroskedasticity." *J. American Stats. Assoc.* 64 (March 1969): 316–23.

Graham, Frank D. *Exchange, Prices and Production in Hyper-Inflation: Germany 1920–1923.* Princeton, N.J.: Princeton Univ. Press, 1930.

Lucas, Robert E., Jr. "Some International Evidence on Output-Inflation Tradeoffs." *A.E.R.* 63 (June 1973): 326–34.

Mills, Frederick C. *The Behavior of Prices.* New York: Arno (for Nat. Bur. Econ. Res.), 1927.

Parks, Richard W. "Inflation and Relative Price Variability." *J.P.E.* 86, no. 1 (February 1978): 79–96.

Phelps, Edmund S. "Introduction: The New Microeconomics in Employment and Inflation Theory." In *Microeconomic Foundations of Employment and Inflation Theory,* by Edmund S. Phelps et al. New York: Norton, 1970.

Sheshinski, Eytan, and Weiss, Yoram. "Inflation and Costs of Price Adjustment." *Rev. Econ. Studies* 44 (June 1977): 287–303.

Statistisches Reichsamt. *Statistisches Jahrbuch für das deutsche Reich.* Berlin: Hobbung, 1921/22 and 1923.

———. *Sonderhefte zur Wirtschaft und Statistic: Zahlen zur Geldenwertung in Deutschland 1914 bis 1923.* Berlin: Hobbung, 1925.

Stigler, George J., and Kindahl, James K. *The Behavior of Industrial Prices.* New York: Columbia Univ. Press (for Nat. Bur. Econ. Res.), 1970.

Vining, Daniel R., and Elwertowski, Thomas C. "The Relationship between Relative Prices and the General Price Level." *A.E.R.* 66 (September 1976): 699–708.

Exchange-Rate Dynamics: An Empirical Investigation

Robert A. Driskill

University of California, Davis

This paper estimates a reduced-form exchange-rate equation whose estimates are used to address questions on exchange-rate overshooting, intermediate-run exchange-rate dynamics, and long-run proportionality relationships between relative money supplies and exchange rates. Based on Swiss–U.S. data from the period 1973–79, the major findings are that following a monetary shock there is short-run exchange-rate overshooting by a factor of about two, and that subsequent exchange-rate adjustments to a new long-run equilibrium take longer than 2 years and exhibit nonmonotonic patterns.

I. Introduction

Much of the recent literature on floating exchange rates has viewed relative money supplies as the basic determinant of exchange rates and relative price levels. While this literature is in agreement about the long-run relationships between money, exchange rates, and price levels, it differs in its predictions about short- and intermediate-run exchange-rate determination.[1] In particular, one strand of this literature, mainly associated with the work of Dornbusch (1976), has emphasized the role of slowly adjusting commodity price levels for short- and intermediate-run exchange-rate dynamics. The most striking implication of the Dornbusch model is that the exchange rate may "overshoot" in the short run: In response to a change in relative

This study stems from my doctoral dissertation. Grateful acknowledgments are due to my primary advisor, Jürg Niehans, and also to Bela Balassa, Jeffrey Frankel, Dale Henderson, George Kanatas, Ed Kane, Jerry Thursby, and an anonymous referee.

[1] For a discussion of these different approaches, see Bilson (1978).

[*Journal of Political Economy*, 1981, vol. 89, no. 2]
© 1981 by The University of Chicago. 0022-3808/81/8902-0011$01.50

money supplies, the exchange rate may immediately change by more than its long-run equilibrium value. Another implication of the model is that, following an initial overshoot, the exchange rate then monotonically approaches its long-run equilibrium value. Both of these implications result from the key assumptions of slowly adjusting commodity prices and perfect capital substitutability and mobility.

Within the theoretical framework of slowly adjusting commodity prices, an alternative to Dornbusch's view of exchange-rate dynamics has developed, an alternative which emphasizes stock/flow interactions and relative-price trade-balance effects. This literature, developed by, among others, Branson (1976), Niehans (1977), and Henderson (1980), alters the strong Dornbusch conclusions about short-run overshooting and monotonic exchange-rate and price-level adjustments to long-run equilibrium, instead permitting short-run undershooting and nonmonotonic exchange-rate and price-level adjustments.

The purpose of this paper is to examine empirically the predictions of these exchange-rate dynamic models against Swiss/U.S. data over the period 1973–77. Specifically, the paper addresses the questions of whether there is short-run exchange-rate overshooting, whether the intermediate-run exchange-rate and price-level adjustments are nonmonotonic, and whether the long-run proportionality relationships between money supplies, price levels, and exchange rates hold.

This is not the first attempt to test a model of exchange-rate determination based on the Dornbusch exchange-rate dynamics approach. Frankel (1979) extends the Dornbusch model by incorporating secular inflation rates and tests it against the deutschmark/dollar rate, and Bilson (1978) provides suggestive evidence about the appropriateness of the Dornbusch approach for the deutschmark/pound rate. Neither test, however, uses an exchange-rate equation rich enough to incorporate trade-balance responses to relative price changes and thus exclude the possibility of more complex, perhaps oscillatory exchange-rate adjustment paths and exchange-rate undershooting.

The major empirical findings from this study are as follows: (1) The exchange rate overshoots in the quarter in which a monetary change takes place by a factor of about two. (2) The exchange-rate adjustment path to full equilibrium is not monotonic, as predicted by the original Dornbusch fixed-output model, but rather exhibits periods of appreciation and depreciation. The price level, however, adjusts monotonically. (3) Purchasing power parity holds in the long run. As noted above, there is wide theoretical agreement that this should be the case, but it has often been difficult to verify empirically. Furthermore, the "long run" is calculated to be approximately 2–3 years.

II. The Reduced-Form Equations

The goal of this section is to motivate the reduced-form exchange-rate and price-level equations which are estimated in the following section. Reduced-form equations are, of course, consistent with a variety of structural models. What is done here is to develop two structural models, the Dornbusch model and a stock/flow model developed by generalizing the Dornbusch model to permit imperfect capital mobility. Both of these structural models impose a priori constraints on the reduced-form parameters and thus can in principle be rejected by the data. Of course, neither structural model may be sufficiently complex, and a "true" structural model may impose few theoretical predictions about a priori reduced-form constraints. This possibility emphasizes the fact that the reduced-form empirical estimates are more general than either structural model.

A. The Dornbusch Model

This subsection recasts the Dornbusch model in discrete time. The model consists of three basic building blocks: a money-market equilibrium condition, a price-level adjustment equation, and an uncovered-interest-arbitrage specification. Assuming that both the domestic and foreign country have identical structural parameters, we can specify money demand as:

$$m_t^d = p_t + \phi y_t - \lambda r_t + v_t, \tag{1}$$

where m is the log of the ratio of the domestic to foreign money supply, p is the log of the ratio of the domestic to foreign price level, y is the log of the ratio of domestic to foreign real income, r is the difference between domestic and foreign interest rates, λ is the interest rate semielasticity of the demand for money in each country, ϕ is the income elasticity of the demand for money in each country, and v_t is a serially uncorrelated random variable with zero mean and variance σ_v^2.

Each country's money supply is assumed to be exogenously controlled by the respective government. Furthermore, all changes in the relative money supply are assumed to be unanticipated; this assumption is important because it implies that this period's relative money supply is also the expected supply for all future periods. Assuming equilibrium obtains each period, we can write the equilibrium condition as:

$$m_t = (m_t)^d = p_t + \phi y_t - \lambda r_t + v_t. \tag{2}$$

In the goods market, relative demand for output depends on relative real income, the relative interest rate, and relative prices and has the following functional form:[2]

$$\log D_t = \gamma y_t - \sigma r_t + \omega(e_t - p_t), \tag{3}$$

where e is the log of the exchange rate and D is demand for domestic output.

The rate of relative inflation, $p_{t+1} - p_t$, is proportional to the log of the ratio of relative demand to relative supply in the goods market:

$$p_{t+1} - p_t = \delta(\log D_t - y_t). \tag{4}$$

Combining equations (2), (3), and (4), we can write the relative price equation as:

$$p_t = a_0 y_{t-1} + a_1 p_{t-1} + a_2 m_{t-1} + a_3 e_{t-1}, \tag{5}$$

where: $a_0 = \delta(1 - \gamma) + \phi/\lambda$, $a_1 = 1 - \delta\sigma/\lambda - \delta\omega$, $a_2 = \delta\sigma/\lambda$, and $a_3 = \delta\omega$.

The final building block of the Dornbusch model is a joint assumption of uncovered interest arbitrage and exchange-rate expectations. The uncovered interest arbitrage assumption can be stated as

$$r_t - x_t = 0, \tag{6}$$

where x_t is the expected change in e from t to $t + 1$. The assumption about x_t is that it is a fraction of the gap between the current and long-run equilibrium exchange rate. Given the assumption that the relative money supply follows a random walk, this implies the following relationship:

$$x_t = \theta(m_t - e_t) + k,\ 0 < \theta < 1,\ k \text{ a constant}. \tag{7}$$

Combining equations (2), (5), (6), and (7), we can derive the following reduced-form equation:

$$e_t = \pi_0 + \pi_1 e_{t-1} + \pi_2 m_t + \pi_3 m_{t-1} + \pi_4 p_{t-1} + \pi_5 y_t + \pi_6 y_{t-1} + \pi_7 z_t, \tag{8}$$

where z_t is a first-order serially correlated random variable and where the π_i's satisfy the following constraints:[3]

[2] Output demand could be specified as a function of the *real* interest rate. In both the Dornbusch and the stock/flow models, this change makes no difference in the qualitative implications derived for the reduced-form equations.

[3] Details are found in Appendix A.

$$\sum_{i=1}^{4} \pi_i = 1 \qquad \pi_3 < 0$$

$$\pi_1 < 0 \qquad \pi_4 < 0$$

$$\pi_2 > 1 \qquad \pi_5 < 0 \qquad \pi_6 < 0.$$

The constraint on the sum of the first four π_i's simply says that purchasing power parity holds in the long run. The constraint that $\pi_2 > 1$ says that there must be short-run overshooting.

B. The Stock/Flow Model

As noted in the Introduction, the stock/flow model is developed by generalizing the Dornbusch model to permit imperfect capital mobility. To do this, a net demand for foreign assets is specified as a linear function of the expected net yield:

$$B_t = \eta(x_t - r_t), \qquad \eta > 0. \tag{9}$$

In addition, a trade-balance equation is specified as a linear function of the log of relative prices and the log of relative real incomes:

$$T_t = \alpha(e_t - p_t) - \beta y_t + u_t, \qquad \alpha, \beta \geq 0, \tag{10}$$

where u_t is a zero-mean, finite-variance, serially uncorrelated random variable.[4]

The foreign-exchange market-clearing equation states that net capital flows equal net trade flows plus all other autonomous flows (assumed constant):

$$\Delta B_t = T_t + A_t, \tag{11}$$

where A_t is a constant. Replacing equation (6) in the Dornbusch model by equation (11), we can derive the following reduced-form exchange-rate equation:

$$e_t = \pi_0' + \pi_1' e_{t-1} + \pi_2' m_t + \pi_3' m_{t-1} + \pi_4' p_{t-1} + \pi_5' y_t + \pi_6' y_{t-1} + \pi_7' z_t'. \tag{12}$$

The π_i''s satisfy the following constraints:[5]

[4] The term T_t could be specified as a function of current and lagged relative prices where the current-price effect is perverse, i.e., creates a J-curve effect. This specification is indistinguishable from (10) in terms of restrictions on reduced-form parameters.

[5] Details are in Appendix A.

$$\sum_{i=1}^{4} \pi_i' = 1 \qquad \pi_3' \gtrless 0$$

$$\pi_1' < 1 \qquad \pi_4' > 0$$

$$\pi_2' > 0 \qquad \pi_5' \gtrless 0 \qquad \pi_6' \gtrless 0.$$

While the purchasing power parity constraint holds just as with the Dornbusch structural model, the other constraints are quite different. In particular, both the lagged exchange-rate and lagged price-level coefficients may be positive in this model. The estimated sign of these coefficients should provide a sharp test between these two models. Also note that π_2' need not be greater than one; this model makes the question of overshooting or undershooting an empirical question.

III. Estimation Results

The reduced-form equations (5), (8), and (12) were estimated for Swiss/U.S. data with two changes. First, after some experimentation with various proxies for Swiss income (data on which are available only on a yearly basis), the income variables y_t and y_{t-1} were dropped. None of the proxies used in the experimentation provided significant coefficients, and the remaining coefficients were hardly affected when income was dropped. Second, two dummy variables were added to the exchange-rate equation: OIL, which takes the value of 1 for December-January-February 1973–74 and zero for all other periods, reflecting the improved attractiveness of dollar-denominated assets following the announcement of the oil embargo and cuts in Arab oil production; and SEAS, which takes the value of 1 every December-January-February and zero all other periods, capturing the pronounced year-end demand for Swiss francs by Swiss firms for end-of-year "window dressing" of their financial statements.[6]

The choice of the Swiss franc/U.S. dollar rate for the period 1973–77 is based on the relative "cleanliness" of the franc float, the substantial variation in the franc-dollar rate, and the independence over this period of the franc and dollar from the "snake."[7]

The estimation uses quarterly average data, where the quarterly data were generated by averaging three monthly end-of-period figures; that is, March-April-May is the average of March, April, and May. This provides a sample size of 19 observations, beginning March

[6] Schiltknecht (1976) discusses this point.

[7] Schiltknecht (1976) indicates that the Swiss have no exchange-rate targets for the Swiss money supply at least for periods longer than a month or two.

1973 and ending November 1977.[8] Even though monthly data would have had the advantage of more observations, the quarterly specification was used since it is more consistent with the theoretical models developed, that is, money-market equilibrium and trade-balance adjustment. While it would be possible to specify partial asset-adjustment schemes and longer trade-balance lags and use the monthly data, this would introduce substantial multicollinearity problems, precluding precise parameter estimates.

Sources and methods of deriving the data are explained in Appendix B. The price index used was the consumer price index, and the monetary aggregate used was M3. Not surprisingly, the results were sensitive to the choice of monetary aggregate: Over the period in question, U.S. M1 demand has not been very stable in comparison with M3 demand.[9] In the presentation of estimated equations, t-statistics are in parentheses beneath the estimated coefficient, and Durbin's H statistic is presented.[10] All estimations were done using White's (1978) SHAZAM regression package.

The estimated relative price equation is:

$$\begin{aligned} p_t &= \underset{(1.90)}{.60} + \underset{(4.58)}{.76}p_{t-1} + \underset{(1.96)}{.37}m_{t-1} - \underset{(.73)}{.03}e_{t-1}, \\ R^2\,(\text{adj.}) &= .96, \qquad \text{Durbin's } H = 1.67. \end{aligned} \tag{13}$$

Hence, $a_1 = .76$, $a_2 = .37$, and $a_3 = -.03$. Both a_1 and a_2 are significant at the .05 level for a one-tailed test, and the hypothesis that $a_1 + a_2 + a_3 = 1$ is not rejected at the .05 level. Since a_3 in (13) is insignificant and the wrong sign, the equation was also estimated with only p_{t-1} and m_{t-1} as regressors.[11] The estimated equation is:

$$\begin{aligned} p_t &= \underset{(1.86)}{.57} + \underset{(4.74)}{.78}p_{t-1} + \underset{(1.89)}{.29}m_{t-1}, \\ R^2\,(\text{adj.}) &= .96, \qquad \text{Durbin's } H = 1.65. \end{aligned} \tag{14}$$

Again, both a_1 and a_2 are significant, and the hypothesis that $a_1 + a_2 = 1$ is not rejected at the .05 level. Finally, the equation was estimated

[8] A regression using end-of-period quarterly data from 1973:II through 1977.III is also reported. More will be said about the choice between average and end-of-period at that point.

[9] See Goldfeld (1976) for evidence on this point.

[10] In regressions with a lagged dependent variable, Durbin's H statistic is the appropriate statistic to test serial correlation of the disturbance terms. For a sample as small as used here, however, the usefulness is diminished.

[11] See Appendix C for a derivation of the conditions under which $a_3 = 0$, but the trade balance is still a function of relative prices. Basically it requires that the trade balance be a very small fraction of total aggregate demand.

with the constraint that $a_1 + a_2 = 1$ imposed:

$$p_t = \underset{(1.84)}{.57} + \underset{(4.47)}{.71}p_{t-1} + \underset{(1.86)}{.29}m_{t-1}. \tag{15}$$

The estimates of the relative price equations, then, are consistent with the model and have good explanatory power.[12]

Preliminary to testing the exchange-rate equation, it is useful to see whether the assumption used in both structural models that the relative money supply follows a random walk, that is, that $m_t = m_{t-1} + x_t$, where x_t is a serially uncorrelated random variable with zero mean, is reasonably consistent with the sample data. To do this, two procedures were used. First, m_t was regressed against a constant term plus the previous lagged values of m_t. A representative result is:

$$m_t = \underset{(.71)}{.15} + \underset{(2.88)}{.99}m_{t-1} - \underset{(-1.08)}{.54}m_{t-2} - \underset{(-.14)}{.08}m_{t-3} + \underset{(1.50)}{.77}m_{t-4} - \underset{(-.70)}{.29}m_{t-5} + \underset{(.79)}{.23}m_{t-6}. \tag{16}$$

Only the coefficient on m_{t-1} is significant, and it is not significantly different from one. The second procedure used a test suggested by Box and Jenkins (1970, p. 291) for randomness in a residual vector from a time-series process. For this test, m_t was regressed on a constant and on m_{t-1} with the following result:

$$m_t = \underset{(.09)}{.01} + \underset{(13.9)}{1.01}m_{t-1}. \tag{17}$$

Residuals from this equation and from an equation in which the constant term and m_{t-1} coefficient were constrained at (0, 1) were then used in the aforementioned test for nonrandomness. Nonrandomness was rejected for both sets of residuals at the 10 percent level. Consequently, the assumption that all changes in the relative money supply are unanticipated seems justified.

The estimate of the exchange-rate equation is:[13]

[12] The relative price equation can be disaggregated by country and written in the following form $P_t^{SW} = d_0 + d_1p_t^{US} + d_2p_{t-1}^{SW} + d_3p_{t-1}^{US} + d_4m_t^{SW} + d_5m_t^{US}$. If the structural equations of Switzerland and the United States are indeed identical, the d_i's should satisfy the following restrictions, $d_1 = 1, d_2 = -d_3, d_4 = -d_5$. This equation was estimated, and the above hypothesis about the d_i's was not rejected at the .05 level, the standard errors, however, were quite large.

[13] As with the price equation, the exchange-rate equation was also estimated with country-disaggregated data to test the assumption of identical structural equations across countries. The equation is $e_t = h_0 + h_1m_t^{SW} + h_2m_t^{US} + h_3m_{t-1}^{SW} + h_4m_{t-1}^{US} + h_5p_{t-1}^{SW} + h_6p_{t-1}^{US} + h_7e_{t-1}$. The coefficient restrictions consistent with identical structural parameters across countries are: $h_1 = -h_2, h_3 = -h_4, h_5 = -h_6$. The hypothesis test of these restrictions was not rejected at the .05 level, the assumption of identical structural parameters cannot be rejected. Furthermore, the point estimates were rather similar to those of eq. (18) and satisfied all the other a priori coefficient restrictions.

$$
\begin{aligned}
e_t = {} & \underset{(-2.82)}{-2.22} + \underset{(3.65)}{.43}e_{t-1} + \underset{(5.73)}{2.37}m_t - \underset{(5.60)}{2.45}m_{t-1} + \underset{(2.23)}{.93}p_{t-1} \\
& + \underset{(7.61)}{.15}\mathrm{OIL} - \underset{(-5.47)}{.06}\mathrm{SEAS};\ \mathrm{RHO} = \underset{(1.37)}{.35}, \\
& R^2\,(\mathrm{adj.}) = .99, \qquad \text{Durbin's } H = .21,
\end{aligned}
\tag{18}
$$

where RHO is the estimated first-order autoregressive parameter of the error term.[14]

The overall explanatory power of the equation is quite good. Furthermore, all coefficients except RHO are significantly different from zero at the .05 level for a one-tailed test. The Durbin H statistic indicates no serial correlation in the disturbances.

As far as a priori constraints on the parameters are concerned, first note that the sum of π_1 through π_4 is 1.28, which is insignificantly different from one at the .05 level. That is, the estimated equation is consistent with the purchasing power parity proportionality prediction about changes in relative money supplies and changes in long-run equilibrium exchange rates. To further test this result, equation (17) was also estimated with this constraint imposed:

$$
\begin{aligned}
e_t = {} & \underset{(-2.89)}{-2.38} + \underset{(4.48)}{.55}e_{t-1} + \underset{(4.92)}{2.30}m_t - \underset{(5.79)}{2.69}m_{t-1} + \underset{(1.91)}{.84}p_{t-1} \\
& + \underset{(7.25)}{.16}\mathrm{OIL} - \underset{(-5.11)}{.06}\mathrm{SEAS};\ \mathrm{RHO} = \underset{(.79)}{.21}.
\end{aligned}
\tag{19}
$$

These coefficients are quite close to the unconstrained ones, thus providing further evidence that the purchasing power parity principle holds.

Both structural models developed in Section II imply that the purchasing power parity constraint should hold. The two models differed sharply, though, on their predictions of the sign of the lagged exchange-rate and lagged price-level coefficients. The estimation results of positive coefficients on these variables are consistent with the stock/flow model but inconsistent with the Dornbusch one. Thus, the data reject the Dornbusch model. Note, though, that the Dornbusch model's major insight about overshooting is verified as an empirical phenomenon.

Two more procedures were carried out to examine further the robustness of the preceding empirical results. First, the exchange-rate equation was estimated using end-of-period rather than average data.

[14] Estimation of (18) was done using a modified Cochrane-Orcutt procedure in the SHAZAM computer program developed by Kenneth White. Grid search and maximum likelihood nonlinear methods produced almost identical estimates.

While using end-of-period data increases the likelihood that any one data point is an outlier, it has the advantage of not possibly confusing anticipated and unanticipated changes in the money supply. The results are:

$$\begin{aligned} e_t = {} & \underset{(-2.66)}{-3.16} + \underset{(3.44)}{.56e_{t-1}} + \underset{(2.44)}{1.69m_t} - \underset{(-3.50)}{2.47m_{t-1}} + \underset{(2.14)}{1.17p_{t-1}} \\ & + \underset{(5.63)}{.21\text{OIL}} - \underset{(-3.44)}{.07\text{SEAS}}; \ \text{RHO} = \underset{(-.50)}{-.14}, \\ & R^2 \text{ (adj.)} = .97, \qquad \text{Durbin's } H = -.34. \end{aligned} \tag{20}$$

The estimated coefficients are significant at the .05 level, have the same signs as those of equation (18), have similar magnitudes, and satisfy the purchasing power parity constraint; indeed, $\Sigma_{i=1}^{4} \pi_i = .94$, which is not significantly different from one at the .05 level.

Second, an in-period dynamic simulation of the exchange-rate equation was performed. This is important because of the presence of a lagged dependent variable and tends to show whether the estimated regression simply picks up the serial correlation in the dependent variable. The forecast error and squared forecast error were then regressed against time to see whether there is a tendency for it to increase with time. The results are as follows:

$$\text{Forecast error} = \underset{(3.83)}{.05} - \underset{(-.14)}{.0002(\text{time})}, \tag{21}$$

$$\text{Forecast error} = \underset{(2.15)}{.003} + \underset{(.22)}{.00003(\text{time})^2}. \tag{22}$$

This provides rather strong support for the empirical results.

IV. Concluding Comments

This paper has addressed those issues in exchange-rate determination raised by the recent work on exchange-rate dynamics. The exchange-rate equation developed was flexible enough to be consistent with, among others, the structural model of Dornbusch (1976) and a stock/flow model incorporating features emphasized by Niehans (1977), Henderson (1979), and others. Both structural models implied that a purchasing power parity a priori constraint on the reduced-form parameters should be satisfied; the estimates did not reject this hypothesis. Other a priori constraints implied by the Dornbusch model were rejected, though. There was, however, empirical verification of Dornbusch's overshooting hypothesis.

To help summarize the empirical findings on these questions, estimated coefficients from equations (15) and (18) are used to calculate an exchange-rate and price-level adjustment path following a one-time monetary increase. To do this, equations (15) and (18) are transformed into final form (suppressing all random disturbance terms):

$$e_t = C_0 + \sum_{i=0}^{\infty} \mu_i m_{t-i}, \qquad C_0 \text{ a constant,} \tag{23}$$

$$P_t = C_0' + \sum_{i=1}^{\infty} \epsilon_i m_{t-i}, \qquad C_0' \text{ a constant,} \tag{24}$$

where the μ_i's and ϵ_i's are as follows:

$$\mu_1 = \pi_2,$$

$$\mu_2 = (\pi_3 + \pi_1\pi_2),$$

$$\mu_3 = (\pi_3 + \pi_1\pi_2)\pi_1 + a_2\pi_4,$$

$$\mu_4 = (\pi_3 + \pi_1\pi_2)\pi_1^2 + a_2\pi_4(\pi_1 + a_1),$$

$$\mu_5 = (\pi_3 + \pi_1\pi_2)\pi_1^3 + a_2\pi_4(\pi_1^2 + a_1\pi_1 + a_1^2),$$

$$\vdots$$

$$\mu_n = (\pi_3 + \pi_1\pi_2)\pi_1^{n-1} + a_2\pi_4(\pi_1^{n-2} + \pi_3^{n-3}a_1 + \pi_1^{n-4}a_1^2 + \ldots + \pi_1 a_1^{n-3} + a_1^{n-2}),$$

$$\epsilon_1 = a_2,$$

$$\epsilon_2 = a_2 a_1,$$

$$\epsilon_3 = a_2 a_1^2,$$

$$\vdots$$

$$\epsilon_n = a_2 a_1^{n-1}.$$

Using these final form coefficients and assuming the initial exchange rate and relative price level are ($e = 0$, $p = 0$), then, following a one-unit unanticipated monetary increase, we can calculate the adjustment path as follows:

$e_0 = 2.30$	$p_0 = 0$	$e_0 - p_0 = 2.30$
$e_1 = .88$	$p_1 = .29$	$e_1 - p_1 = .59$
$e_2 = .34$	$p_2 = .50$	$e_2 - p_2 = -.16$
$e_3 = .21$	$p_3 = .65$	$e_3 - p_3 = -.44$
$e_4 = .27$	$p_4 = .75$	$e_4 - p_4 = -.48$
$e_5 = .39$	$p_5 = .82$	$e_5 - p_5 = -.43$
$e_6 = .48$	$p_6 = .87$	$e_6 - p_6 = -.39$
$e_7 = .54$	$p_7 = .91$	$e_7 - p_7 = -.37$
$e_8 = .61$	$p_8 = .94$	$e_8 - p_8 = -.33$
$e_9 = .67$	$p_9 = .96$	$e_9 - p_9 = -.29$
$e_{10} = .73$	$p_{10} = .97$	$e_{10} - p_{10} = -.24.$

The exchange rate initially depreciates to 2.30, then appreciates for three quarters to 0.21, and then depreciates again. After 11 quarters, about three-quarters of the exchange-rate adjustment has taken place, and 97 percent of the price adjustment has occurred. Notice that deviations from purchasing power parity are positive for only the first two quarters, after that remaining negative. This means that exports are stimulated and imports suppressed because of increased relative prices for only one quarter, and then the effects are reversed.[15] Note also that the exchange-rate path to long-run equilibrium is, as anticipated, nonmonotonic, showing periods of both appreciation and depreciation.

To keep these empirical results in perspective, some of the more serious limitations should be noted. First, the exchange-rate equation tested was based on a bilateral model, while in reality the $n - 1$ independent exchange rates of an n-country world are functions of exogenous variables of all n countries. The equations estimated here, then, are subject to omitted-variables specification bias. If both the correlation of these omitted variables with the included variables and the expected sign of the coefficient of these omitted variables can be determined, then the bias in the estimated coefficients can be determined. At present, not even the expected sign of the omitted coefficients has been determined.[16]

Second, the equation estimated is a reduced form from whose

[15] Note that the cumulative trade balance over the adjustment path must be zero, since there is no net accumulation of foreign assets.

[16] See Berner et al. (1975) for an attempt to develop an n-country model.

estimates structural parameter values cannot be retrieved. While the reduced-form estimates seem plausible, they would be more credible if they were shown to be consistent with existing estimates of an underlying structural model's parameters. Still, the results provide general support for the monetary approach to exchange-rate determination and specific support for the importance of both slow commodity-price adjustments and current-account relative-price effects within that approach.[17]

Appendix A

I. *Derivation of Reduced-Form Constraints from the Dornbusch Model*

Combining equation (7) with (6) and (2), we get:

$$e_t = m_t(1 + 1/\theta\lambda) + p_t(-1/\theta\lambda) + y_t(-\phi/\theta\lambda). \tag{A1}$$

Substituting (5) for p_t we have:

$$\begin{aligned} e_t = {} & e_{t-1}(-a_3/\theta\lambda) + m_t(1 + 1/\theta\lambda) + m_{t-1}(-a_2/\theta\lambda) + p_{t-1}(-a_1/\theta\lambda) \\ & + y_t(-\phi/\theta\lambda) + y_{t-1}(-a_0/\theta\lambda). \end{aligned} \tag{A2}$$

II. *Derivation of Reduced-Form Constraints from the Stock/Flow Model*

Replacing equation (6) of the Dornbusch model with equation (11) we get the following difference equation:

$$\begin{aligned} e_t = {} & e_{t-1}\left[\frac{\eta\theta + a_3(\alpha - \eta/\lambda)}{\eta\theta + \alpha}\right] + m_t\left(\frac{\eta\theta + \eta/\lambda}{\eta\theta + \alpha}\right) \\ & + m_{t-1}\left(\frac{-\eta\theta - \eta/\lambda - a_2\eta/\lambda + a_2\alpha}{\eta\theta + \alpha}\right) + p_{t-1}\left[\frac{(1 - a_1)\eta/\lambda + \alpha a_1}{\eta\theta + \alpha}\right] \\ & + y_t\left(\frac{\beta}{\eta\theta + \alpha}\right) + y_{t-1}\left\{\frac{a_0(\alpha - \eta/\lambda)}{\eta\theta + \alpha}\right\}. \end{aligned} \tag{A3}$$

Note that the coefficient on m_t may be less than one, the coefficient of e_t may be positive, and the coefficient on p_{t-1} must be positive.

Appendix B

I. *Switzerland*

1. *Exchange rate.*—End-of-period data are end-of-month single observations. Source: OECD Main Economic Indicators, Paris, various issues. Quarterly data are averages of daily noon New York rates. Source: *Federal Reserve Bulletin*, Washington, various issues.

2. *Money supply.*—The M3 is taken from International Financial Statistics of the IMF. It appears to include some demand deposits denominated in

[17] Again, see Bilson (1978) for a discussion of how the Dornbusch approach fits within the general monetary framework.

foreign currencies. From figures supplied by Peter Buomberger of the Swiss National Bank, it appears that this component is a small percentage of total M3 (less than 4 percent) and fairly stable. These figures were only for 22 nonconsecutive months, which precluded using them for estimation purposes.

3. *Consumer price index.*—Source: OECD Main Economic Indicators. These are monthly end-of-period figures.

II. *United States*

1. *Money supply.*—Source: *Federal Reserve Bulletin.* Seasonally unadjusted data were used for the quarterly estimations.

2. *Consumer price index.*—Source: OECD Main Economic Indicators.

Appendix C

This Appendix shows that a_3 in the price equation may be close to zero and the trade balance still a function of relative prices if the trade balance is a relatively small fraction of aggregate demand for a country's output. Let D be the ratio of domestic to foreign aggregate demand, E and E^* domestic and foreign expenditure, respectively, that is, $C + I + G$, and P relative prices between the two countries' outputs. The question is what implicit assumption makes $(d \log D)/(d \log P)$ close to zero. Now,

$$D = \frac{E + T}{E^* - T}, \tag{C1}$$

where T is the trade balance. Hence,

$$\begin{aligned} \frac{d \log D}{d \log P} &= \frac{1}{(E + T)} \frac{d(E + T)}{d \log P} - \frac{1 d(E^* - T)}{(E^* - T) d \ln P} \\ &= \frac{P}{(E + T)} \frac{d(E + T)}{dP} - \frac{P}{(E^* - T)} \frac{D(E^* - T)}{dP}. \end{aligned} \tag{C2}$$

Now define the trade balance elasticity as $PT'/T = \epsilon$. Assuming that $dE/dP = dE^*/dP = 0$ and that $E = E^*$, we have

$$\frac{d \log D}{d \log P} = \frac{2\epsilon}{(E/T) - (T/E)}. \tag{C3}$$

Hence, $(d \log D)/(d \log P)$ goes to zero as E/T goes to infinity, despite the size of ϵ. Note that this does *not* say that trade balance responsiveness to relative prices does not affect exchange rates; it only does not affect aggregate demand. What is more important for exchange-rate determination are the relative sizes of trade flows and capital flows.

References

Berner, Richard B.; Clark, P.; Howe, Howard; Kwack, Sung Y.; and Stevens, G. "Simultaneous Determination of the U.S. Balance of Payments and Exchange Rates: An Exploratory Report." International Finance Discussion Paper no. 59, Board of Governors of the Federal Reserve, Washington, February 1975.

Bilson, John F. O. "The Current Experience with Floating Exchange Rates: An Appraisal of the Monetary Approach." *A.E.R. Papers and Proc.* 68 (May 1978): 392–97.

Box, George E. P., and Jenkins, Gwilym M. *Time Series Analysis, Forecasting and Control.* San Francisco: Holden-Day, 1970.

Branson, William H. "Asset Markets and Relative Prices in Exchange Rate Determination." Seminar Paper no. 66, Institute for International Economic Studies, Stockholm, 1976.

Dornbusch, Rudiger. "Expectations and Exchange Rate Dynamics." *J.P.E.* 84, no. 6 (December 1976): 1161–76.

Frankel, Jeffrey A. "On the Mark: A Theory of Floating Exchange Rates Based on Real Interest Differentials." *A.E.R.* 69 (September 1979): 610–22.

Goldfeld, Stephen M. "The Case of the Missing Money." *Brookings Papers Econ. Activity*, no. 3 (1976), pp. 683–730.

Henderson, Dale. "The Dynamic Effects of Exchange Market Intervention Policy: Two Extreme Views and a Synthesis." In *Economics of Flexible Exchange Rates*, edited by Helmut Frisch and Gerhard Schwödiauer (*Kredit und Kapital*, vol. 6 [suppl.], 1980).

Niehans, Jurg. "Exchange Rate Dynamics with Stock/Flow Interaction." *J.P.E.* 85, no. 6 (December 1977): 1245–57.

Schiltknecht, K. "Monetary Policy under Flexible Exchange Rates: The Swiss Case." Mimeographed. Swiss National Bank, 1976.

White, Kenneth J. "A General Computer Program for Econometric Methods—SHAZAM." *Econometrica* 46 (January 1978): 239–40.

The Family as an Incomplete Annuities Market

Laurence J. Kotlikoff

Yale University and National Bureau of Economic Research

Avia Spivak

Ben Gurion University

Families can self-insure against uncertain dates of death through implicit or explicit agreements with respect to consumption and interfamily transfers. Interfamily transfers need have nothing to do with altruistic feelings; they may simply reflect risk-sharing behavior of completely selfish family members. Although family annuity markets are incomplete, even small families can substitute by more than 70 percent for perfect market annuities. Given adverse selection and transaction costs, family risk pooling may be preferred to public market annuities. In the absence of public annuities, these risk-sharing arrangements provide powerful incentives for marriage and family formation.

The institution of the family provides individuals with risk-sharing opportunities which may not otherwise be available. Within the family there is a degree of trust and a level of information which alleviates three key problems in the provision of insurance by markets open to the general public, namely, moral hazard, adverse selection, and deception. In addition, provision of insurance within the family may entail smaller transaction costs than arise in the purchase of insurance on the open market. There are a number of important risks for which

We are grateful for financial support from the Foundation for Research in Economic Education and the National Bureau of Economic Research. Any opinions expressed are solely our own. We wish to thank Finis Welch, Joe Ostroy, Bryan Ellickson, John McCall, Steven Shavell, John Riley, Jon Skinner, and Gary Galles for helpful discussions.

Journal of Political Economy, 1981, vol 89, no 2]
© 1981 by The University of Chicago 0022-3808/81/8902-0006$01 50

the "public" market problems of moral hazard, adverse selection, and deception are especially severe. The risk of loss of job or earnings because of changes in the pattern of demand or partial disability is one example. Here the ability of the public market to determine the extent to which the individual actually suffered an earnings loss or is simply lying about his backache is highly questionable. Other examples are the risk of bankruptcy and the default risk on personal loans. Many family practices in dealing with these types of risks can be explained as implicit insurance contracts made ex ante by completely selfish family members. Love and affection may be important for the enforcement of some of these implicit contracts, but they need not be their sole or even chief determinant. Healthy brother A's support for disabled brother B may simply be the quid pro quo for brother B's past implicit promise to support A if A became disabled instead of B.

The existing economics literature on marriage and the family (including Schultz [1974] and Becker, Landes, and Michael [1977]) has not, to our knowledge, explicitly considered the family's role in providing insurance to family members.

This paper is concerned with family provision of insurance against the risk of running out of consumption resources because of greater than average longevity. The problem is how fast to consume over time when one does not know how long one will continue to live. Too much consumption when young may mean relative poverty later on if one lives "too long"; alternatively, excessive frugality when young involves the risk of dying without ever having satisfied one's hunger. A complete annuity market permits an individual to hedge this uncertainty of the date of death by exchanging his initial resources for a stream of payments that continue as long as the individual survives. We demonstrate here that implicit risk-sharing arrangements within marriage and the family can substitute to a large extent for the purchase of annuities in public markets. Since the number of family members involved in the risk pooling is generally small, these family risk-sharing arrangements constitute an incomplete annuities market. However, our findings suggest that even small families can substitute by more than 70 percent for a complete annuity market in pooling the risk of death. When the economic structure of society is sufficiently developed to sustain organized public insurance markets, implicit risk pooling within an incomplete family annuity market may well be preferred to public purchase of annuities because of adverse selection and transaction costs.[1] When organized insurance markets do not

[1] The transaction costs we have in mind here include the time costs involved in negotiating individual specific annuity contracts. As we demonstrate in the text, each individual's optimal annuity contract depends on his rate of time preference, his degree

exist, the analysis here indicates that implicit risk-sharing arrangements can provide powerful economic incentives for marriage and family formation.

Throughout the paper individuals are assumed to be completely selfish; that is, they obtain utility only from their own consumption. One implication of this approach is that voluntary transfers from children to parents or bequests and gifts from parents to children need have nothing at all to do with altruistic feelings; rather, they may simply reflect risk-sharing behavior of completely selfish individuals. While altruism per se is not required, some level of mutual trust and honesty *is* required since elements of these arrangements are not legally enforceable.

This paper is divided into four sections, the first of which describes optimal consumption behavior for a single individual in both the presence and absence of a complete annuity market. The welfare gains from access to a complete and fair annuity market are calculated for the case of the iso-elastic utility function. This welfare gain is, in turn, decomposed into income and substitution effects. This decomposition suggests that an important component of the gains from access to complete or incomplete annuity markets is the desirability of substituting future for current consumption.

Section II develops the theoretical argument for Pareto-efficient implicit family annuity contracts and explores potential welfare gains arising from these arrangements.[2] Although the complexity of the calculations precluded analysis of large families, quantitative results for families of two and three persons are presented. The analysis considers cases in which family members both do and do not have identical survival probabilities (i.e., are of similar and dissimilar ages and sexes). This framework permits us to ask whether marriage between individuals with similar survival probabilities is more efficient than marriage between individuals with dissimilar survival probabilities.

Optimal family annuity contracts involve agreements on the consumption path of each family member as well as a commitment on the part of each member to name the other members as sole heirs in his estate. Section III discusses the problems of enforcing both aspects of these agreements. Section IV summarizes the paper and suggests areas for future research.

---

of risk aversion, and his survival probabilities. Some individuals may prefer a constant annuity stream, others an increasing or decreasing stream of annuity payments.

[2] Kotlikoff and Spivak (1979) present a proof that family annuity contracting converges to a complete annuities market as the number of family members increases.

I. A Single Person's Consumption Plans with and without Fair Annuities[3]

In the absence of an annuity market, a single individual's consumption choice problem is to maximize his expected utility, equation (1), from current and future consumption subject to the budget constraint, equation (2):

$$EU = \sum_{t=0}^{D} P_t U(C_t), \tag{1}$$

$$\sum_{t=0}^{D} C_t R^{-t} = W_0. \tag{2}$$

The P_t's of equation (1) are probabilities of surviving from age zero through age t; P_0 equals one. The term D is the maximum longevity. For simplicity, we assume the utility function is separable in consumption (C_t) over time. In (2) R, the discount factor, is one plus the interest rate. The initial wealth of the individual is W_0; we ignore possible streams of future labor earnings or inheritances.[4]

The budget constraint written in equation (2) is identical to the budget constraint that would arise in a certainty world in which individuals never died before age D. While individuals will, on the average, die prior to age D, equation (2) reflects the nonzero probability that an individual will live through age D; that is, equation (2) is the relevant budget constraint because the individual may actually live through age D, in which case his realized present value of consumption cannot exceed his budget.

Let us now assume that the single person is free to purchase actuarially fair annuities in a complete public annuities market. The budget constraint in this case is

$$\sum_{t=0}^{D} P_t C_t R^{-t} = W_0. \tag{3}$$

[3] Yaari (1965) is the pioneering paper on this subject. Sheshinski and Weiss (1981) provide an illuminating discussion on the interaction of annuities and social insurance. Barro and Friedman (1977) provide an analysis of the risks of the uncertainty of the date of death.

[4] The gains from access to an annuities market are greatest when the individual has all his resources up front. This assumption, then, dramatizes the demand for annuities; but dropping this assumption would not alter the theoretical point that families can substitute for annuity markets. For the sake of completeness one can think of the individuals described in this paper as having received all their resource streams prior to their current age. In the no-annuity, no-family world involuntary bequests can be thought of as being collected by the government and redistributed to individuals at their birth.

In contrast with (2), (3) requires only an equality between the expected present value of consumption and initial wealth. The single individual now chooses his optimal consumption path by maximizing (1) subject to (3); he then exchanges his initial wealth W_0 with the insurance company in return for its promise to pay out the C_t stream as long as the person continues to live.

The P_tR^{-t}'s in (3) may be thought of as prices. Since each of the P_t's in (3), except P_0 which equals unity, is less than one, the consumption choice in the case of a fair annuity market is equivalent to the consumption choice without an annuity market but with lower prices of future consumption. Obviously, access to a fair annuity market increases utility by expanding the budget frontier; it also alters the optimal consumption path because of the income and substitution effects resulting from the lower prices of future consumption.

The iso-elastic utility function (4) is convenient for assessing the potential gains from access to a fair public annuities market as well as the gains from family annuity arrangements:

$$EU = \sum_{t=0}^{D} P_t \frac{C_t^{1-\gamma}}{1-\gamma} \alpha^t. \tag{4}$$

In (4), γ is the constant relative risk-aversion parameter, and α is the time preference parameter. By considering different values of γ we indicate for this family of utility functions how the gains from annuities and family arrangements depend on the specification of tastes.

In the no-annuities case maximization of (4) subject to (2) leads to the consumption plan, (5):

$$C_t = \frac{W_0(R\alpha)^{t/\gamma}P_t^{1/\gamma}}{\Sigma_{j=0}^{D}R^{j(1-\gamma)/\gamma}\alpha^{j/\gamma}P_j^{1/\gamma}}; \tag{5}$$

In the case of fair annuities, maximizing (4) subject to (3) leads to

$$C_t = \frac{W_0(R\alpha)^{t/\gamma}}{\Sigma_{j=0}^{D}R^{j(1-\gamma)/\gamma}\alpha^{j/\gamma}P_j}. \tag{6}$$

Figure 1 compares equations (5) and (6) for the case $R = \alpha = 1$. The ability to trade in a fair annuities market may raise or lower initial consumption, depending on whether γ is less than or greater than unity (fig. 1). Intuitively, the higher the degree of risk aversion, γ, the greater the concern for running out of money because of excessive longevity and, hence, the lower the initial consumption. At γ equal to infinity, equation (5) dictates equal consumption in each period.

Plugging (5) or (6) into (4), we arrive at two indirect utility functions for the no-annuity and annuity cases with initial wealth, the interest

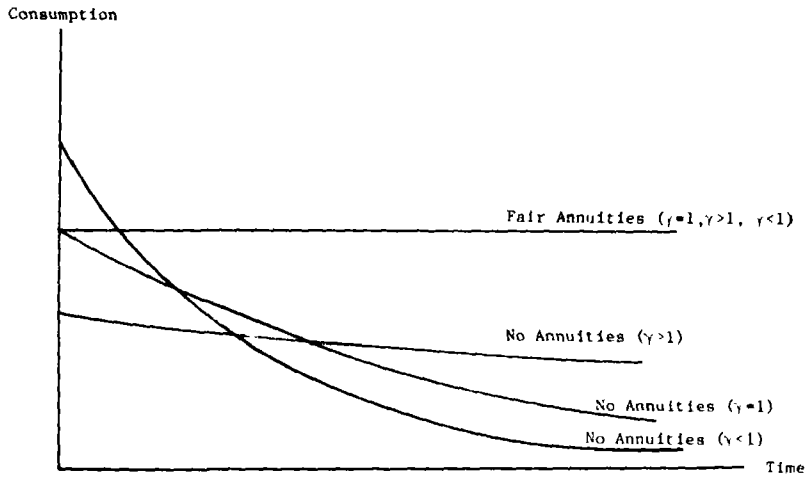


FIG. 1.—Consumption paths with and without fair annuities

rate, and survival probabilities as arguments. These functions are presented in equations (7) and (8), respectively,

$$H_0(W_0) = \frac{1}{1-\gamma} W_0^{1-\gamma} \left[\sum_{j=0}^{D} \alpha^{j/\gamma} R^{j(1-\gamma)/\gamma} P_j^{1/\gamma} \right]^{\gamma}, \tag{7}$$

$$V_0^*(W_0) = \frac{1}{1-\gamma} W_0^{1-\gamma} \left[\sum_{j=0}^{D} \alpha^{j/\gamma} R^{j(1-\gamma)/\gamma} P_j \right]^{\gamma}. \tag{8}$$

The increase in utility resulting from access to fair annuities can be measured in terms of dollars. Equation (9) solves for the value of M, which represents the percentage increment in a single person's initial wealth required, in the absence of an annuity market, to leave him as well off as he would be with no additional wealth but with access to an annuities market:

$$H_0(MW_0) = V_0^*(W_0). \tag{9}$$

For the iso-elastic utility function this calculation is independent of the initial level of wealth. Table 1 reports values of M for different ages and levels of risk aversion using both male and female survival probabilities. Friend and Blume (1975) estimate the degree of relative risk aversion from individual portfolio choices. They conclude that risk aversion, on average, exceeds unity. We present our results for risk-aversion coefficients of 0.75, 1.25, and 1.75, a range that we feel

TABLE 1

PERCENTAGE INCREASE IN INITIAL WEALTH REQUIRED TO OBTAIN FAIR ANNUITIES UTILITY LEVEL

Age	Relative Risk Aversion (γ)	Males	Females
30	.75	24.5	18.5
55	.75	46.9	34.4
75	75	71 2	63.0
90	75	99.8	100.2
30	1.25	30.3	22.7
55	1.25	59.2	43.4
75	1.25	97.0	85.3
90	1.25	152.6	152.9
30	1.75	34.7	26.1
55	1.75	68 9	50.7
75	1.75	119.1	104.6
90	1.75	199 1	199.4

NOTE — Throughout table α = 99 and R = 1 01.

encompasses reality. The survival probabilities used in this and all subsequent calculations are actuarial estimates from the Social Security Administration.[5] Maximum longevity is taken to be 120 throughout the paper.

Table 1 indicates that the utility gain measured in dollars from access to an annuities market can be quite large. For a relative risk-aversion parameter value of 0.75, the gain to a 55-year-old male is equivalent to a 46.90 percent increase in his initial wealth. The utility gain is age dependent; for $\gamma = 0.75$, the 30-year-old male's gain is 24.46 percent, while the 90-year-old male's gain is 99.81 percent. Annuities are less important to young people because a large fraction of their lifetime utility from consumption is fairly certain due to their lower mortality probabilities in the immediate future. Higher levels of risk aversion naturally increase the gains from access to an annuities market. The male-female differences in the table reflect the higher male age-specific mortality rates. The calculation is somewhat sensitive to the choice of α and R. Raising the interest rate to 5 percent while holding α constant increases the age 55 wealth-equivalent factor from 46.90 to 55.57 for the case of $\gamma = 0.75$. The 90-year-old wealth equivalent is increased from 99.81 to 115.34.

[5] We use the low mortality male and female probabilities reported on pp. 17 and 19 of the Social Security Administration Actuarial Study no. 62 (see U.S. Department of Health, Education, and Welfare 1966).

Income, Substitution Effects, and Unintended Bequests

Without access to an annuity market a single, nonaltruistic individual will always die prior to consuming all his wealth and, accordingly, will make involuntary bequests. The level of these unintended bequests can be quite large. From equation (5) we calculated the consumption path as well as the corresponding wealth path for the no-annuity case. By multiplying the probability of dying at each age times the wealth at each age and discounting back to the initial age, the present expected value of these unintended bequests can be computed. For $\gamma = 0.75$, $R = 1.01$, and $\alpha = .99$, the present expected value of unintended bequests represents 24.47 percent of initial wealth for a single male aged 55. This number means that a 55-year-old male with no annuity market will, on average, fail to consume about one-quarter of his wealth because he is risk averse. Increasing the risk-aversion coefficient to 1.75 raises the ratio of present unintended bequests to initial wealth to 0.3583. These large unintended bequests occur despite a fairly rapid rate of consumption. Current mortality probabilities dictate a fairly rapid rate of consumption even for high levels of risk aversion. For $\gamma = 1.75$, a single male who survives to age 85 consumes at age 85 less than a third of his age 55 consumption level.[6]

The homothetic property of the iso-elastic utility function permits a decomposition of the utility gains from fair annuities into income and substitution effects. Suppose a fair insurance company approached a single, 55-year-old ($\gamma = 0.75$) male and offered to pay him 24.47 percent of his initial wealth in exchange for his naming the insurance company as his heir. The single male would take the 24.47 percent gain and, because of homotheticity, consume it according to his original no-annuity consumption path. This additional wealth would give rise to an additional $.2447 \times .2447$ in present expected bequests. By letting the insurance company also pay for this second round of expected but involuntary bequests as well as further rounds, the insurance company ends up paying $32.40 = .2447/(1 - .2447)$ percent of the single individual's initial wealth. This 32.40 percent figure represents the utility gain from the pure income effect. In this scenario the individual continues to consume at the no-annuity set of prices. Since the total gain from being able to purchase fair annuities and thus face lower prices for future consumption is 46.90 percent, the income effect represents 69.08 percent and the substitution effect 30.92 percent of the total gain. Hence, the ability to alter the age

[6] The ratio of consumption at age 75 to consumption at age 55 is 0.62. When risk aversion equals 0.75, the ratio of consumption at age 85 to consumption at age 55 equals 0.06; it is 0.33 at age 75.

consumption profile is an important part of the total welfare gain from annuities.

II. The Family as an Incomplete Annuities Market

Decisions by family members concerning consumption expenditures and interfamily transfers may reflect implicit though incomplete annuity contracts. In the case of marriage both individuals commonly agree to pool their resources while both marriage partners are alive and to name each other as the major, if not the sole, beneficiary in their wills. For each partner the risk of living too long is somewhat hedged by the other partner's potential death; if one partner lives to be very old, there is a high probability that his (or her) spouse has already died leaving him a bequest to help finance his consumption. While each spouse gains simply from the exchange of wills, the two can further increase their expected utilities by agreeing on a joint consumption path that takes into account each spouse's expected bequest to the other. The importance of joint consumption planning is highlighted in the case of an implicit contract between a parent and a child. Here the parent implicitly promises to name the child in his will in exchange for the child's implicit promise to care for the parent if the parent lives too long. Although the child may have zero probability of dying while the parent is still alive, both can gain because the child agrees to share consumption resources with the parent.

This view of bequest and consumption arrangements within marriage as an incomplete annuity market becomes intuitive when one contemplates increasing the number of members in the family. To simplify the issue, let us assume that all individuals within the family have identical survival probabilities and that they enter this multiperson family with identical resources. In the limit as the family (or "tribe") gets large, the consumption path of an individual within the tribe converges to the path a single individual would choose in a complete and actuarially fair annuities market (Kotlikoff and Spivak 1979).

Quantitative Analysis of Family Risk Pooling

In the case of two family members the frontier of efficient marriage contracts is obtained as the solution to the following recursive dynamic programming problem:

$$\begin{aligned} V_{t-1}(W_{t-1}) = \max_{W_t, C^H_{t-1}, C^S_{t-1} \geq 0,\ t = 1, \ldots, T} & [u^H(C^H_{t-1}) + \theta u^S(C^S_{t-1}) + \alpha P_{t/t-1} Q_{t/t-1} V_t(W_t) \\ & + \alpha P_{t/t-1}(1 - Q_{t/t-1}) H_t(W_t) + \theta\alpha Q_{t/t-1}(1 - P_{t/t-1}) S_t(W_t)], \end{aligned} \tag{10}$$

subject to

$$W_t/R + C^H_{t-1} + C^S_{t-1} = W_{t-1}, \tag{11}$$

where

$$V_T(W_t) = \max_{C^H_T, C^S_T} u^H(C^H_T) + \theta u^S(C^S_T).$$

In (10) $V_t(W_t)$ is the period t maximum-weighted expected utility of the two family members with joint wealth W_t. In the expression the letters H and S denote the two family members, C^H_t and C^S_t are the consumptions of the two, u^H and u^S are their utility functions, $P_{t/t-1}$ and $Q_{t/t-1}$ are their respective period t survival probabilities conditional upon surviving through period $t - 1$, and $H_t(W_t)$ and $S_t(W_t)$ are the maximum expected utilities for each member if he or she alone survives to period t. These expressions are obtained from equation (7) by replacing W_0 with W_t and applying the appropriate probabilities. The term θ is the differential weight applied to member S's expected utility.

The first two terms on the right-hand side of (10) represent utility from certain period $t - 1$ consumption. The third term is the family's expected period t utility multiplied by the probability that both members survive to period t. The last two terms represent expected utilities when one member dies and the other survives.

The Appendix presents an algorithm to solve (10). The algorithm for solving the three-family-member maximization problem is available from the authors.[7]

The Gains from Family Annuity Contracts

The solution to (10) permits a comparison of consumption paths and utility levels of married people with those of single persons, assuming throughout that there is no public annuities market. Both spouses are assumed to have identical iso-elastic utility functions in the sense of the same degrees of risk aversion and rates of time preference.

The shape of consumption paths for married couples while they are both alive may differ from that of single individuals for two reasons. First, even if the two spouses have identical survival probabilities, the reduction in risk within the marriage rate acts like a reduction in the price of future consumption. If the relative risk-aversion parameter γ exceeds (is less than) one, the identical survival probabilities marriage profile will start above (below) the single person's profile. For γ equal

[7] In the three-member family we maximize a weighted sum of the three members' expected utility taking all survival contingencies into account. If one of the three dies first, the other two jointly inherit the remaining wealth and consume according to the optimal two-person plan.

to unity the profiles are identical. In terms of figure 1, the consumption profiles for married persons lie between the no-annuity and complete annuity profiles. The second reason for different consumption profiles for married people relative to single individuals is possible differences in spousal survival probabilities. Higher survival probabilities act like lower rates of time preferences. When an old man marries a young woman the slope of the optimal marriage consumption profile reflects the survival probabilities of both the old husband and the young wife. The two spouses compromise with respect to the rate at which they eat up their joint wealth while they are both alive. The old husband would prefer to eat up the wealth more rapidly, and the young wife would prefer to consume at a slower rate. The formula for each spouse's consumption when married takes both spouses' survival probabilities into account as well as the relative spousal utility weights. To our knowledge empirical studies of consumption and savings at the household level have not considered this point—that the time preference rate for a household may depend on the age-sex composition of the household.

Table 2 reports the gains from marriage as well as three-person polygamy among individuals who have identical survival probabilities and identical initial endowments and who are weighted equally in the contract. The marriage and three-person polygamy gains are calculated as the percentage increase in a single person's initial wealth needed to make him as well off as he would be in the marriage or polygamous relationship. The table also reports the dollar gain as a fraction of the table 1 total dollar gain from complete and fair annuities. Since utility is concave in wealth, the dollar gain from these family contracts as a fraction of the dollar gain from fair annuities is smaller than the actual utility gain from these contracts as a fraction of the utility gain from fair annuities. Table 2 also reports this latter fraction.[8]

The figures in table 2 indicate that marriage can offer substantial risk-pooling opportunities. For a 55-year-old man using male survival probabilities, pooling risk through marriage is equivalent to about a 20 percent increase in his wealth had he stayed single. The gains from marriage increase as one becomes older since the risks incurred are much greater as one ages. At age 75 marriage is equivalent to increasing one's wealth by 30 percent when risk aversion is 1.25. Death-risk-pooling through marriage can be quite important even at young ages. The table reports gains from 11.7 to 13.6 percent at age 30 using the male probabilities.

[8] This fraction is calculated as $[(1 + m)^{1-\gamma} - 1]/[(1 + a)^{1-\gamma} - 1]$, where m is the fractional wealth equivalent gain from marriage, and a is the fractional wealth equivalent gain from fair annuities.

TABLE 2

THE ANNUITY GAINS FROM MARRIAGE AND THREE-PERSON POLYGAMY

		Marriage			Three-Person Polygamy		
Age	Risk Aversion	Dollar Marriage Gain (%)	Dollar Marriage Gain/Dollar Annuity Gain	Utility Marriage Gain/Utility Annuity Gain	Dollar Polygamy Gain (%)	Dollar Polygamy Gain/Dollar Annuity Gain	Utility Polygamy Gain/Utility Annuity Gain
30	.75	11.7	.478	.499	15.8	.645	.665
55	.75	20.0	.426	.461	28.0	.597	.632
75	.75	25.4	.357	.405	37.2	.522	.571
90	.75	28.2	283	.339	43.1	.432	.496
30	1.25	13.0	.429	470	18.0	.594	.635
55	1.25	22.3	.377	.446	32.1	.542	.612
75	1.25	30.1	.310	.419	45.2	.466	.571
90	1.25	37.1	.243	.367	58.2	.381	.524
30	1.75	13.6	.392	.456	19.3	.556	.619
55	1.75	23.5	.341	.451	34.5	.501	.613
75	1.75	33.2	.279	.449	50.7	.426	.596
90	1.75	43.0	.216	.420	68.7	.345	.579

NOTE.—The table uses male mortality probabilities $R = 1.01$ and $\alpha = .99$

Marriage can also close much of the utility gap between no annuities and complete annuities. For example, for a 55-year-old with risk aversion of 0.75, marriage substitutes 46.10 percent for complete and fair annuities. Marriage is a better substitute for fair annuities at younger ages because at younger ages the probability that both spouses will die simultaneously is quite small relative to the probability that one spouse will die before the other. In addition, there appears to be an interaction between age and the degree of risk aversion, making marriage a better substitute for fair annuities at young ages when risk aversion is low and at old ages when risk aversion is high.

Over a wide range of ages and parameter values, three people appear to be capable of capturing about 60 percent of the gains from fair annuities. While the complexity of the calculations precluded considering a four-person arrangement, we can conjecture using table 2 how well four people would do together. In the case of a 55-year-old male with risk aversion of 0.75, adding one marriage partner is equivalent to a 20 percent increase in his wealth had he remained single. The marginal dollar gain from adding a third person (table 2) is 8.04 percent. If the marginal dollar gain fell at a constant rate in this range, the fourth person would add 8.04 × (8.04/20.0) = 3.23 percent.[9] By adding 3.23 to 28.04, we can roughly calculate the extent to which four people can close the utility gap. The procedure suggests that four people can substitute by 70 percent for a fair annuities market.

Diminishing returns to risk pooling appear, then, to set in at a fairly rapid rate. In this example two people substitute by 46 percent, three people by 63 percent, and four people by over 70 percent for full insurance.

Table 3 considers incomplete annuity arrangements between two parents and one child and between one parent and two children. In both cases we assume equal consumption by all family members but permit the initial wealth of the child or children to vary. All individuals are assigned the male survival probabilities; the children are age 30 and the parents age 55. In the case of two parents with one child, if the child has an initial wealth of $35,000 and the parents have an initial wealth of $20,000, entering into an equal consumption–will-swapping arrangement is equivalent to a 32 percent increase in wealth for each parent and a 10.6 percent increase for each child. For the parent this arrangement captures 71.2 percent of the utility gain from

[9] This is probably a lower bound estimate for the contribution of the fourth person; the marginal dollar gain cannot fall at a constant 40 percent rate forever, because if it did the total dollar gains would, in the limit, not sum up to 46.9 percent, the full annuity gain of table 1. Presumably the marginal dollar gain falls at a decreasing rate, and 3.23 percent probably underestimates the fourth person's marginal contribution.

TABLE 3

GAINS FROM INCOMPLETE ANNUITY ARRANGEMENTS IN THE FAMILY

Initial Wealth of Each Child ($)	Two Parents with One Child: Dollar Gain to Parent (%)	Two Parents with One Child: Dollar Gain to Child (%)	Two Children with One Parent: Dollar Gain to Parent (%)	Two Children with One Parent: Dollar Gain to Child (%)
25,000	14.4	34.2	2.3	24.8
30,000	23.2	20.4	16.9	18.9
35,000	32.0	10.6	31.5	14.6
40,000	40.8	3.2	46.1	11.5

NOTE.—The calculations assume equal consumption by all family members. Initial wealth of parent or parents is $20,000. $R = 1.01$, $\alpha = .99$, and $\gamma = 0.75$.

full annuities; for the child the arrangement substitutes by 45.4 percent for full annuities. The last two columns of table 3 present the case of two children contracting with one parent. When each child contributes $35,000, the gain to the parent is 31.5 percent, while each 30-year-old child enjoys a 14.6 percent gain relative to consuming as a single person. The numerical differences in the table for the two different types of families reflect, on the one hand, different monetary contributions of parents relative to children and, on the other hand, differences in the rate at which resources are consumed when all family members are alive. Resources are consumed at a slower rate in the two-children–one-parent case than in the one-child–two-parent case, since each individual's survival probabilities are given equal weight in determining the optimal rate of consumption.

Is Marrying People of Similar Ages More Efficient?

Suppose one had to decide how to pair up four people, two who are old and two who are young. Is it more efficient to marry the old people together and the young people together than it is to mix ages? Intuitively, marrying two 90-year-olds together and two 20-year-olds together leaves a large chance that both 90-year-olds will die in the immediate future, and resources that they have failed to consume will be lost to the 20-year-olds who, on average, will still be alive.[10] The countervailing argument against mixed-age marriages is that mixing

[10] We assume here that any involuntary bequests that arise from the simultaneous death of both marriage partners or from the death of a surviving spouse are not inherited by any of these four individuals. Again the government can be thought of as collecting these residual bequests and distributing them each year to the newborn. We thank Finis Welch for a helpful discussion on this section.

ages involves greater risk to one of the two partners; the utility cost of this greater risk may exceed the utility gain from the increase in expected resources arising in mixed marriages.[11]

We investigated potential efficiency gains from mixed marriages between two 55-year-olds and two 30-year-olds, where each individual was risk averse at the 0.75 level. The 55-year-olds were assigned male survival probabilities, while the 30-year-olds were assigned female survival probabilities. When risk aversion equals 0.75, weights of 1.7 for the old person yield utility levels for both old and young which exceed those in the old-old, young-young marriages of table 2. The additional dollar gain to the old person from this weighted marriage with the young person is 3.1 percent; the added gain to the young person is 1.6 percent.

These additional gains from mixed-age marriages require, however, a fairly skewed distribution of consumption within the marriage. For this example the young-old weighting scheme that dominates old-old, young-young coupling involves the older spouse's consuming about 86 percent more than the younger spouse while they are both alive. If it is too costly to negotiate such an arrangement within the marriage or if the type of consumption (e.g., housing) within marriage is nonexcludable, then equal consumption marriages of individuals with similar survival probabilities (of similar ages) will be the rule rather than the exception. Of course, we have been discussing here marriages in which each spouse has the same initial dowry. The old-young marriages can dominate old-old, young-young marriages even under an equal consumption arrangement provided the dowry of the young spouse sufficiently exceeds that of the old spouse.

III. Enforcement with and without Altruism

In the absence of altruism would family members voluntarily maintain these implicit contracts as family members age? The answer is

[11] To see this consider an old-young marriage in which the young person promises to consume less than the old person in the state of nature in which both spouses survive. Suppose that this promise to the old person of higher consumption in the "both survive" state is large enough to exactly compensate the old person for the loss in expected utility from the state in which his spouse dies but he survives. The old person's expected utility from this latter "bequest" state is lower when he marries someone young, rather than someone old, because the probability of the young person actually dying is smaller. While the old person is by assumption no worse off in this compensated old-young marriage, the young person could be worse off than if he had married someone young. By entering into the compensated old-young marriage, the young person reduces his payoff in the both survive state while leaving the payoff in the bequest state unchanged. He also increases the probability of the bequest state and lowers the probability of the both survive state. Although expected consumption for the young spouse rises, the spreading of the payoffs may lower expected utility, depending on the young spouse's degree of risk aversion.

that there always are ways of structuring payments to individuals within the family so that each individual at each moment in time has a selfish interest in maintaining the original implicit contract. An equal consumption marriage contract between two individuals with the same survival probabilities and the same initial endowment is a good first example. If each spouse maintains control over his own wealth while both spouses are alive and consumes at the same rate as the other spouse, then each will separately have an incentive to continue the contract at each point in time. A similar type of individual control can be maintained in family arrangements; rather than have the parents use up all their resources before the children begin contributing to their support, the children can contribute each period in return for that period's expected parental bequest. This scenario of parents' maintaining control over their wealth until the very end, as enforcement leverage over their children, may partly explain the limited use of gifts as a tax-saving intergenerational transfer device.

The proof of this proposition is immediate from equation (10). Given their initial endowments at time $t - 1$, W^H_{t-1} and W^S_{t-1}, family members choose a value θ^* such that the contract to consume the contingent plan $[C^H_{t-1}(\theta^*), C^S_{t-1}(\theta^*), C^H_t(\theta^*), C^S_t(\theta^*), \ldots, C^H_D(\theta^*), C^S_D(\theta^*)]$ is in the core at time $t - 1$. The consumption plan at time t $[C^H_t(\theta^*), C^S_t(\theta^*), \ldots, C^H_D(\theta^*), C^S_D(\theta^*)]$ represents the period t Pareto-efficient contract corresponding to the initially chosen utility weight θ^*. This plan is in the core for a set S_t of individual endowments of family members in period t, W^H_t and W^S_t, which satisfy $W^H_t + W^S_t = W_t$. To insure that the initial contract remains a core allocation for each family member, side payments are made at time $t - 1$ when consumption in period $t - 1$ occurs. The side payments leave the period t individual endowments in set S_t. Since the initial contract $[C^H_{t-1}(\theta^*), C^S_{t-1}(\theta^*), C^H_t(\theta^*), C^S_t(\theta^*), \ldots, C^H_D(\theta^*), C^S_D(\theta^*)]$ is in the core, each selfish family member will have a personal incentive to make or accept these side payments.

There are two additional questions of enforcement to consider. One problem is that a spouse may covertly name a third party as beneficiary in his will in exchange for the same commitment by the third party or in exchange for a particular service. A second type of cheating may occur when one or both spouses covertly consume in excess of the consumption levels dictated by an optimal implicit marriage contract; while each spouse may correctly believe that he or she is the beneficiary in the other spouse's will, each may try to take advantage of the other by increasing his own consumption and thus reducing the potential bequest available to the other spouse.

These two types of cheating will be more problematic for implicit incomplete annuity agreements between friends or relatives who are

physically separated. The consumption cheating scenario can be modeled as a Nash equilibrium in which each partner chooses his consumption path by taking the other partner's consumption path and potential bequest path as given. Resources are consumed at a faster rate in the Nash equilibrium as each partner fails to consider how his consumption will diminish his expected bequest and thus the expected utility of his partner.

Using male survival probabilities we calculated for two 55-year-olds the dollar equivalent utility gain from engaging in a Nash consumption-cheating partnership. The gains in the Nash equilibrium proved to be almost identical to those in the more efficient marriage contract. For levels of risk aversion of 0.75, 1.25, and 1.75, the percentage dollar increments are, respectively, 19.9, 22.2, and 23.5. While the rate of consumption is faster in the Nash equilibrium, it is not much faster than in the marriage contract. Intuitively cheating by overconsuming is fine provided one's partner actually dies; but if one's partner survives, the early excessive consumption will require relative deprivation later on. Apparently this latter consideration dominates the former, leaving utility in the cheating equilibrium at essentially the same level as under a marriage contract. These examples suggest that consumption cheating does not represent a substantial impediment to consumption–risk-sharing arrangements.

Another means of enforcing these implicit contracts is simply altruism. All of our calculations have involved maximizing a weighted sum of individual family members' utilities. If, however, each family member is altruistic toward each other and each weights each family member's utility from consumption in the same way, then all family members would unanimously agree on the utility maximand. The calculations we have presented can, therefore, be thought of as resulting from the maximization of an agreed-upon altruistic family utility function. Since all family members agree on the maximand, there is no problem of enforcement.

IV. Summary and Conclusion

This paper has demonstrated that consumption and bequest-sharing arrangements within marriage and larger families can substitute to a large extent for complete and fair annuity markets. In the absence of such public markets, individuals have strong economic incentives to establish relationships which provide risk-mitigating opportunities. Within marriages and families there is a degree of trust, information, and love which aids in the enforcement of risk-sharing agreements. Our calculations indicate that pooling the risk of death can be an important economic incentive for family formation; the paper also

suggests that the current instability in family arrangements may, to some extent, reflect recent growth in pension and social security public annuities. The methodological approach of this paper can be applied to the study of family insurance against other types of risks. Of chief interest are those types of risks that are handled very poorly by anonymous public markets. Disability insurance and insurance against earnings losses are good examples.

Our approach has been to compare family insurance with perfect insurance. It would seem worthwhile to compare family insurance with public market insurance where the market insurance is subject to adverse selection and moral hazard problems and family insurance is not. Realistic specification of the degree of adverse selection and moral hazard may indicate that family insurance dominates public market insurance even in small families.

Finally, the paper suggests the empirical difficulty of determining whether intergenerational transfers reflect altruism or simply risk-mitigating arrangements of essentially selfish individuals in the absence of perfect insurance markets. Distinguishing between the selfish and altruistic models is fundamental to a number of major economic questions, including the impact of the social security system on national saving and the effectiveness of fiscal policy.[12]

Appendix

Computational Algorithm for the Two-Family-Members Dynamic Risk-pooling Problem

This Appendix indicates the algorithm used to solve the two-family-members dynamic programming problem, copied here as equation (A1). The algorithm for the case of three family members is similar to that for two members and is available from the authors. While we consider the iso-elastic family of utility functions, our algorithm can be applied to any homothetic utility function.

$$\begin{aligned} V_{t-1}(W_{t-1}) = \max_{W_t, C^H_{t-1}, C^S_{t-1} \geq 0,\, t = T, \ldots, 1} \{ & u^H(C^H_{t-1}) + \theta u^S(C^S_{t-1}) + \alpha P_{t|t-1} Q_{t|t-1} V_t(W_t) \\ & + \alpha P_{t|t-1}(1 - Q_{t|t-1}) H_t(W_t) + \theta \alpha Q_{t|t-1}(1 - P_{t|t-1}) S_t(W_t) \}, \end{aligned} \tag{A1}$$

subject to

$$W_t/R + C^H_{t-1} + C^S_{t-1} = W_{t-1}. \tag{A2}$$

Again, the letters H and S correspond to the two family members with respective conditional survival probabilities $P_{t|t-1}$ and $Q_{t|t-1}$. The expression W_t is joint family wealth, θ is the weighting factor, and $H_t(W_t)$ and $S_t(W_t)$ are the expected utility levels for each family member if he alone survives to period t.

[12] See Barro 1974.

Optimal values for C_t^H and C_t^S are found recursively starting at period T and proceeding to period 0. We demonstrate that $V_t(W_t)$ may be written in the form:

$$V_t(W_t) = v_t \frac{W_t^{1-\gamma}}{1-\gamma}, \tag{A3}$$

where v_t is a constant. We also show that total family consumption, C_t, is given by

$$C_{t-1} = W_{t-1} \frac{v_t^{1/\gamma}}{v_t^{1/\gamma} + (\alpha K_t R)^{1/\gamma} R^{-1}}, \tag{A4}$$

where K_t is another constant. Given total family consumption, consumption of the two members is

$$C_{t-1}^H = \frac{C_{t-1}}{1+\theta^{1/\gamma}}, C_{t-1}^S = C_{t-1} \frac{\theta^{1/\gamma}}{1+\theta^{1/\gamma}}. \tag{A5}$$

We demonstrate that K_t is a function of v_t and that v_{t-1} is a function of K_t. Starting then at the initial value for K_t, K_{T+1}, we can compute v_T; v_T in turn gives K_T, which in turn gives v_{T-1}. Proceeding in this fashion to period zero we compute the entire sequence of v_t's and K_t's. These values can then be used in equation (A4) to compute the ratio of consumption to wealth at each period. These ratios together with an initial level of wealth plus (A2) and (A5) generate the optimal consumption path. The homotheticity of the utility function permits us to calculate recursively the shape of the consumption path independently of the initial level of wealth.

Starting with period T the maximization problem for equation (A1) is

$$V_T(W_T) = \max \frac{1}{1-\gamma}(C_T^H)^{1-\gamma} + \theta \frac{1}{1-\gamma}(C_T^S)^{1-\gamma}$$
$$\text{s.t.} \quad C_T^H + C_T^S \leq W_T, \quad C_T^H, C_T^S \geq 0.$$

Solving this maximization and computing the indirect utility function for V_T, we have

$$V_T(W_T) = v_T \frac{1}{1-\gamma} W_T^{1-\gamma}, \text{ where } v_T = (1+\theta^{1/\gamma})^\gamma, \tag{A6}$$

$$C_T^H = W_T \frac{1}{1+\theta^{1/\gamma}}, C_T^S = W_T \frac{\theta^{1/\gamma}}{1+\theta^{1/\gamma}}. \tag{A7}$$

For $t < T$, (A1) for the iso-elastic case is written as

$$\begin{aligned} V_{t-1}(W_{t-1}) = \max_{C_{t-1}^H, C_{t-1}^S} & \frac{1}{1-\gamma}(C_{t-1}^H)^{1-\gamma} + \frac{\theta}{1-\gamma}(C_{t-1}^S)^{1-\gamma} \\ & + \alpha \frac{P_t}{P_{t-1}} \frac{Q_t}{Q_{t-1}} v_t \frac{1}{1-\gamma} W_t^{1-\gamma} + \alpha \frac{P_t}{P_{t-1}}\left(1 - \frac{Q_t}{Q_{t-1}}\right) h_t \frac{1}{1-\gamma} W_t^{1-\gamma} \\ & + \theta\alpha\left(1 - \frac{P_t}{P_{t-1}}\right) \frac{Q_t}{Q_{t-1}} s_t \frac{1}{1-\gamma} W_t^{1-\gamma} \\ \text{s.t.} \quad & C_{t-1}^H + C_{t-1}^S + \frac{W_t}{R} = W_{t-1}. \end{aligned} \tag{A8}$$

In going from (A1) to (A8) we use the fact that $H_t(W_t) = h_t[W_t^{1-\gamma}/(1-\gamma)]$ and $S_t(W_t) = s_t[W_t^{1-\gamma}/(1-\gamma)]$ for the iso-elastic utility function. The values for h_t and s_t are implicitly defined as the bracketed term in equation (7) in the text with $j = 0$ corresponding to time t and with each family member's survival probabilities from time t substituting for P_j.

It is easy to see from (A8) that for given total family consumption, C_t, C_t^H and C_t^S will always satisfy (A5). Hence we may rewrite (A8) as

$$V_{t-1}(W_{t-1}) = \max_{C_{t-1}} v_T \frac{1}{1-\gamma} C_{t-1}^{1-\gamma} \tag{A9}$$

$$+ \alpha \frac{W_t^{1-\gamma}}{1-\gamma}\left\{\frac{P_t}{P_{t-1}}\frac{Q_t}{Q_{t-1}} v_t + \frac{P_t}{P_{t-1}}\left(1 - \frac{Q_t}{Q_{t-1}}\right)h_t + \theta\left(1 - \frac{P_t}{P_{t-1}}\right)\frac{Q_t}{Q_{t-1}} s_t\right\}.$$

Denoting the term in brackets by K_t we now have

$$V_{t-1}(W_{t-1}) = \max_{C_{t-1}} v_T \frac{1}{1-\gamma} C_{t-1}^{1-\gamma} + \alpha \frac{1}{1-\gamma} W_t^{1-\gamma} K_t \tag{A10}$$

$$\text{s.t.} \quad C_{t-1} + \frac{W_t}{R} = W_{t-1}.$$

Maximizing (A10) and computing the indirect utility functions yields

$$v_{t-1} = [v_T^{1/\gamma} + (\alpha R K_t)^{1/\gamma} R^{-1}]^\gamma, \tag{A11}$$

$$C_{t-1} = W_{t-1} \frac{v_T^{1/\gamma}}{v_T^{1/\gamma} + (\alpha K_t R)^{1/\gamma} R^{-1}}. \tag{A12}$$

References

Barro, Robert J. "Are Government Bonds Net Wealth?" *J.P.E.* 82, no. 6 (November/December 1974): 1095–1170.

Barro, Robert J., and Friedman, James W. "On Uncertain Lifetimes." *J.P.E.* 85, no. 4 (August 1977): 843–49.

Becker, Gary S.; Landes, Elisabeth M.; and Michael, Robert T. "An Economic Analysis of Marital Instability." *J.P.E.* 85, no. 6 (December 1977): 1141–87.

Friend, Irwin, and Blume, Marshall E. "The Demand for Risky Assets." *A.E.R.* 65 (December 1975): 900–922.

Kotlikoff, Laurence J., and Spivak, Avia. "The Family as an Incomplete Annuities Market." Working Paper no. 362, Nat. Bur. Econ. Res., June 1979.

Schultz, Theodore W., ed. *Economics of the Family: Marriage, Children, and Human Capital.* Chicago: Univ. Chicago Press (for Nat. Bur. Econ. Res.), 1974.

Sheshinski, Eytan, and Weiss, Yoram. "Uncertainty and Optimal Social Security." *Q.J.E.* (1981), in press.

U.S. Department of Health, Education, and Welfare, Social Security Administration. *United States Population Projection for OASDHI Cost Estimates.* Actuarial Study no. 62. Washington: U.S. Department of HEW, Social Security Administration, December 1966.

Yaari, Menahem E. "Uncertain Lifetime, Life Insurance, and the Theory of the Consumer." *Rev. Econ. Studies* 32 (April 1965): 137–50.

An Economic Theory of Self-Control

Richard H. Thaler

Cornell University

H. M. Shefrin

University of Santa Clara

The concept of self-control is incorporated in a theory of individual intertemporal choice by modeling the individual as an organization. The individual at a point in time is assumed to be both a farsighted *planner* and a myopic *doer*. The resulting conflict is seen to be fundamentally similar to the agency conflict between the owners and managers of a firm. Both individuals and firms use the same techniques to mitigate the problems which the conflicts create. This paper stresses the implications of this agency model and discusses as applications the effect of pensions on saving, saving and the timing of income flows, and individual discount rates.

For many years Christmas clubs paid no interest. Members deposited money each week but could only withdraw the money on December 1. The clubs were very popular, although they seemed to be dominated by simply depositing money in an interest-bearing savings account.

Passbook loans do still exist. These loans allow an individual with $5,000 in a savings account earning 5 percent interest to borrow at 9 percent using the balance as collateral, instead of at 10 percent with

The authorship of this paper was fully collaborative. The order of the names on this paper and its companion (Shefrin and Thaler 1980) corresponds to our division of labor. This paper was written while Thaler was a visiting scholar at the National Bureau of Economic Research at Stanford, California. While there he received financial support from the Kaiser Family Foundation and the Robert Wood Johnson Foundation. Helpful comments on earlier drafts were provided by George Ainslee, James Buchanan, Tom Russell, and numerous other friends and colleagues. We thank them all.

[*Journal of Political Economy*, 1981, vol. 89, no. 2]
© 1981 by The University of Chicago. 0022-3808/81/8902-0008$01.50

no collateral. Obviously the individual could simply withdraw the money from his savings account at an (opportunity) cost of only 5 percent.

Smoking clinics are a new and thriving business. A smoking clinic will help people who want to stop smoking—for a fee of several hundred dollars.

What does economic theory have to say about these institutions? George J. Stigler provides the following analysis:

> One can of course explain the participation in a Christmas fund by introducing another item of preference: a desire of people to protect themselves against a future lack of willpower. . . . If we stopped the analysis with this explanation, we would turn utility into a tautology: a reason, we would be saying, can always be found for whatever we observe a man to do. In order to preserve the predictive power of the utility theory, we must continue our Christmas fund analysis as follows. The foregone cost of putting money in a Christmas fund is the interest one could earn by putting the same money in a savings account. If interest rates on savings accounts rise, the cost of buying protection against a loss of willpower rises and less of it ought to be bought. . . . [Stigler 1966, p. 57]

We agree with the remarks above and therefore offer a model that says more about these institutions than the fact that the demand for their services will be negatively related to price. We do so by proposing a simple extension of orthodox models, using orthodox tools, that permits such behavior to be viewed as rational. This rationalization is based on an analysis of the technology of self-control using the theory of agency rather than reliance on ad hoc explanations in which transaction costs, taxes, and income effects play a major role. Our model can predict, in a nontautological way, the circumstances in which these kinds of behavior will be observed. In particular, our new theory of intertemporal choice has important implications for theories of saving behavior. In the last section of the paper we discuss some of these implications and offer empirical evidence in support of our ideas.

I. The Model

> The idea of self-control is paradoxical unless it is assumed that the psyche contains more than one energy system, and that these energy systems have some degree

of independence from each other. [DONALD MCINTOSH 1969]

Why individuals would impose constraints on their future behavior is a problem that has received attention from economists since Strotz's (1955–56) classic paper.[1] Strotz and those who have followed him (Pollak 1968; Blackorby et al. 1973; Peleg and Yaari 1973; Hammond 1976; and Yaari 1977) have analyzed the phenomenon as one of *changing tastes*. In Strotz's formulation a conflict occurs between today's preferences and tomorrow's preferences if the discount function used today is not exponential with a constant exponent.

Our framework differs from the changing-tastes literature in that we model man as having two sets of preferences that are in conflict at a single point in time. This idea is certainly not new. Adam Smith used a two-self model much like ours in his *Theory of Moral Sentiments* (1759). More recently Schelling (1960, 1978) and Buchanan (1975) have recognized the importance of simultaneous conflict in understanding self-control problems. Outside economics the idea is commonplace, with the writings of Freud (1958) and Berlin (1969) deserving special mention.

Nonetheless, to the best of our knowledge our work is the first systematic, formal treatment of a two-self economic man. We have adopted a two-self model because, as McIntosh says above, the notion of self-control is paradoxical without it. Furthermore, we utilize an organizational analogy that leads to both insights into human behavior and a rich explanatory model. We will briefly describe the model here. We do so in order to make explicit the nature of our two-self conceptualization. The model also leads to specific predictions about behavior, which are discussed in Section III below. Further details are available in our more formal companion paper (Shefrin and Thaler 1980).

Our model is cast in discrete time. Consider an individual with a fixed income stream $y = (y_1, y_2, \ldots, y_T)$. Think of period T as retirement and let $y_T = 0$. Let the individual choose a nonnegative level of consumption c_t in t. Call $c = (c_1, c_2, \ldots, c_T)$ a consumption plan. The conflict between short-run and long-run preferences is introduced by viewing the individual as an organization. At any point in time the organization consists of a planner and a doer.[2] The planner is concerned with lifetime utility, while the doer exists only for one period and is completely selfish, or myopic. The period t doer is assumed to have direct control over period t consumption rate c_t. The

[1] Two important contributions by noneconomists are Ainslee (1975) and Elster (1977).

[2] These terms originated in an early draft of Thaler's (1980).

doer's utility function is given by $Z_t(\cdot)$; Z_t is taken to be independent of all components of c except c_t.[3] Furthermore, suppose initially that Z_t is strictly increasing and concave in c_t.

In our model the planner does not actually consume but, rather, derives utility from the consumption of the doers. Therefore the planner's utility function is given by $V(Z_1, Z_2, \ldots, Z_T)$. Observe that a plan which maximizes V subject to the present value budget constraint, $\sum_{t=1}^{T} c_t \leq \sum_{t=1}^{T} y_t = Y$, is considered optimal from the planner's point of view. However, without some method to control the doer's actions, this plan cannot be implemented. Indeed, under the assumptions above, the actual consumption stream chosen would have total lifetime income consumed during the first period, when the period-one doer would borrow $Y - y_1$ on the "perfect" capital market. In order to prevent this from occurring, the planner requires some psychic technology capable of affecting the doer's behavior. Two main techniques are available for this: (1) The doer can be given *discretion* in which case either his *preferences* must be modified or his *incentives* must be altered, or (2) the doer's set of choices may instead be limited by imposing *rules* that change the constraints the doer faces.

We begin by analyzing the case in which no rules are used. We refer to this case as *pure discretion*. While we do not believe that the pure discretion case is empirically important (most people appear to use at least some kinds of rules), it provides a useful foundation for our model, to which rules are easily added. Furthermore, since this case corresponds closely to that usually considered in economics, it highlights the differences between our model and the standard framework.

Recall that $Z_t(\cdot)$ was assumed to be unbounded. We now specify that Z_t depends on a preference modification parameter θ_t selected by the planner. The choice of θ_t allows the planner to alter Z_t such that it possesses an internal maximum. By appropriately selecting θ_t, any desired c_t may be obtained; however, the lower the desired c_t is, the more modification will be required. Furthermore, $\partial Z_t/\partial\theta_t$ is negative; that is, modification reduces short-run utility and is therefore costly. Finally, we assume that the marginal cost of modification increases with θ. Thus successive reductions in c_t require increasing reductions in Z_t.

If the planner could exercise direct control over the choice of a consumption plan, θ_t would be set equal to zero for all t (because modification is costly) and c would be selected to maximize V subject to the budget constraint. Since under pure discretion this is assumed to

[3] The extreme assumption that doers do not care at all about past or future doers is adopted just for expositional simplicity. Other arguments of Z_t could easily be added.

be infeasible, the planner must instead choose $\theta = (\theta_1, \theta_2, \ldots, \theta_T)$ to maximize V. (The solution to the problem is described in Shefrin and Thaler [1980].) Essentially, modification is increased until the marginal utility derived from additional consumption in retirement (period T) equals the marginal loss in utility in earlier periods due to modification.

In the more general case when both rules and preference modification are permitted, the planner may also alter the budget constraint facing each doer. This allows the planner to reduce c_t without incurring modification costs. However, since available rules are imperfect (see next section), the planner will have to trade off modification costs with the opportunity costs associated with using second-best-type rules. Rules are formally incorporated into the planner-doer model in Shefrin and Thaler (1980).

II. Techniques to Reduce Conflicts in Individuals and Organizations

We have characterized self-control as an internal conflict resembling the principal-agent conflict between the owner and manager of a firm (see Ross 1973; Jensen and Meckling 1976). In this section we describe the actual techniques used by individuals and firms to mitigate these conflicts. The techniques fall into the two categories highlighted in the model: rules and incentives. We provide many illustrations of the methods individuals use because these are in essence part of our model. When individuals use rules it is impossible to characterize their behavior simply with first-order conditions. The limits on the kinds of rules which individuals will find feasible lead to the specific predictions about saving behavior discussed in Section III. In this section we also compare the techniques individuals use with those used in firms. The close correspondence we find lends intuitive support to our principal-agent model.

A. *Methods to Alter Incentives*

Individuals use three basic techniques to alter the doer's incentives. First, the doer's preferences can be modified directly. Some individuals consider saving a good in and of itself.[4] In this case doer myopia does not inhibit saving. Second, inputs to a saving or dieting program can be explicitly monitored via weekly budgets or calorie counting

[4] This idea has been suggested by Scitovsky (1976). The importance of norms in controlling individual behavior is also stressed heavily by Adam Smith (1759, p. 326) and by Irving Fisher (1930).

(customers of diet clinics and credit counselors are advised to do this). Simply keeping track seems to act as a tax on any behavior which the planner views as deviant. Third, incentives can be explicitly altered: Alcoholics take the drug Antabuse which makes them ill if they take a drink; academics agree to give a paper at a conference to provide a proximate incentive to write it.

Firms use the same three methods. First, profit-sharing plans are quite popular even though they offer only trivial financial incentives to all but the highest executives. We believe firms adopt them because they help create an atmosphere in which the employees' preferences are more similar to those of the owner. Second, firms monitor departmental inputs through cost accounting and then tie compensation to these input measures. Third, departmental profits are measured and used as performance measures for managers.

B. Methods to Alter Opportunities: Rules

If the costs of monitoring and persuasion are high, individuals will resort to rules that restrict the doer's opportunities. In the extreme, all doer discretion can be eliminated using what Strotz referred to as the strategy of precommitment. Such behavior is rational in our model if the rule can approximate the choices that the planner would select. Market precommitment institutions are observed, as we would predict, in such areas as saving and dieting. For example, people pay to go to "fat farms" which essentially are resorts that promise not to feed their customers.

Less extreme rules can limit the *range* of doer discretion, usually through the use of self-imposed rules of thumb. In the savings context several such rules appear to be commonly used. These rules alter the budget constraint faced by the doer in much the same way as credit limits imposed by lenders do.[5] A simple first departure from pure discretion is a ban on borrowing, the so-called debt ethic. A somewhat weaker rule which seems common is to prohibit borrowing except for specific purchases, like houses and automobiles. Another rule of thumb is a prohibition on dissaving combined with limits on borrowing. Using this rule of thumb, a person might borrow and lend simultaneously in spite of a substantial difference in the interest rates, as in the case of the passbook loan. The loan allows him to transfer consumption across time periods while it provides a regimented repayment scheme.

[5] While it is difficult to document the extent to which these precise rules are used, 85 percent of Cagan's (1965) sample of Consumers Union members reported using one of these or similar rules to determine monthly saving.

Rules can also eliminate discretion over a specific *class* of decisions for which the conflict is particularly acute. Dieters try not to keep cheesecake in the refrigerator and will refuse invitations to lavish dinner parties; problem gamblers avoid Las Vegas. Also, many smokers pay more for their cigarettes by buying them by the pack instead of the carton—it helps enforce a self-imposed ration such as one pack a day.

Again, the same types of rules are observed in organizations.[6] "Pure" rules are observed most often in bureaucratic organizations because the costs of monitoring output are so high. Rules that limit managerial discretion over a particular range are frequently in the form of guidelines (e.g., a plant manager can adapt any investment that exceeds some stated rate of return). Similarly, rules may prohibit discretion over a specific class of decisions; loan officers, for example, might need approval for loans to relatives or friends.

We wish to make three other points about internal rules of thumb. First, it is useful to consider these rules as learned as much as chosen. Rules like the debt ethic are learned from parents and other models, which suggests that there will be differences in the use of rules depending on social class, education, and age. Second, rules of thumb are likely to become habits. By establishing a routine, the doer decision process can be avoided. Third, to the extent that the rules do become habits, there will be rigidities built into the individual's behavior.[7] The implications of these observations are discussed in the next section.

III. Implications

We now turn to a discussion of the implications of the planner-doer model. What predictions about behavior can be made with our model that are inconsistent with the standard model? Some of these predictions are obvious. We predict that people will rationally choose to impose constraints on their own behavior. Furthermore, we predict that such precommitments will occur primarily for those goods whose benefits and costs occur at different dates. We present here some less obvious predictions based on our model.

Because our framework is richer than the standard theory, vari-

[6] For a discussion of firms imposing constraints on their future financial policies, see Myers (1977).

[7] Habits can be formally introduced in two stages. At first stage Z_t can also be made a function of θ_s, $s < t$, to reflect the fact that self-control at the early dates renders self-control at later dates less costly. At the second stage, Z_t can also be parameterized on the adopted rule. In this case the function Z_t could exhibit an internal maximum in c_t when $\theta_t = 0$. It is interesting to note that modification would now be required in order to break the rule. This would tend to explain miserliness, for instance.

ables that the standard theory treats as irrelevant differences in form we model as differences in substance. In Sections IIIA and IIIB we consider how differences in the form of payment (holding the level constant) affect saving decisions, and in Section IIIC we present variables other than borrowing and lending rates that determine individual marginal rates of time preference.

A. *Pensions and Saving*

Consider two identical individuals with the same total income and wealth. Assume that they both save some fraction s of their income. Now give one a mandatory pension plan that forces him to save $p < s$. What will happen to total saving? Though it is difficult to get a specific prediction from the standard model (see Feldstein 1977), a first-order prediction would be that total saving is unaffected. Other forms of saving should fall by the approximate amount of the pension.

Our model has a different prediction: The pension plan produces saving at no psychic cost. Modification costs occur only when saving is voluntarily withheld. This is what we call "discretionary saving." Thus since the marginal cost of saving is lower at the old saving level, we expect total saving to go up. (In other words, the offset in other saving will be less than the size of the pension.)

Furthermore, in two cases the offset will be essentially zero: (1) The individual uses a saving rule such as "save s percent of disposable income" which is not changed when the pension is introduced (total saving increases by $[1 - s]py_t$), and (2) the individual uses discretion but treats saving as a good for its own sake rather than as a transfer to future consumption. In this case discretionary saving and retirement saving are not perfect substitutes as in the standard model; in fact, their cross-elasticities of demand could be zero.[8]

The effect of pensions on saving has been investigated with individual data by Cagan (1965), Katona (1965), and Munnell (1974, 1976). All obtained similar results. Cagan used a sample of Consumers Union members. He found that for those members with a mandatory pension plan, other saving actually was higher than for those without pensions (see Cagan 1965, p. 21). Munnell (1974) replicated Cagan's study using the same data source. She used a different measure of saving, replaced before-tax income with after-tax income, and restricted her analysis to a subset of observations she thought to be more reliable. She then regressed the nonpension saving to income ratio on several variables including a pension dummy. Her basic result

[8] A formal treatment of the effect of pensions on saving appears in Shefrin and Thaler (1980).

was that the pension had no effect on other saving (i.e., a zero offset). The coefficient of the pension dummy was never significant. Its highest t value occurred for the 55–65 age group ($t = 1.2$), for which those with pensions saved 3 percent less than those without pensions. Though other explanations (such as selectivity bias) have been offered for these results, we find that they lend support to our model.

B. Saving and the Timing of Income Flows

The importance of current disposable income (as opposed to permanent income) yields another prediction from our model that differs from the standard theory. Consider two identical individuals, S and B; S receives a salary of $12,000 per year paid in 12 monthly installments of $1,000, while B receives a salary of $10,000 per year paid in monthly installments plus a guaranteed bonus of $2,000 paid in March each year. Standard theories of saving behavior would predict that these two individuals would save the same amount. Our model predicts that, on average, B will save more.

Although we know of no test of this hypothesis,[9] there is one bit of circumstantial evidence. In Japan, where there is a very high saving rate, bonus schemes are quite common. We think this is no coincidence. A test would be possible, given the right data. We predict that individuals who are paid a portion of their salary via a lump-sum bonus will have higher saving rates than those who receive their compensation in a smooth pattern. How does this follow from the model?

We have characterized saving behavior primarily as a set of self-imposed rules of thumb and externally enforced saving plans. For an individual like B, those rules and plans will be based on his regular monthly income. Contributions to pension plans, payments on whole life insurance policies, mortgage payments, and so forth must be made on a regular basis. Furthermore, most individuals prefer to have their monthly inflows and outflows roughly balance.[10] For B to act like S, he could deposit the bonus in the bank and draw it down gradually during the year (as if his salary were $12,000), or he could borrow the $2,000 over the course of the year and repay the loan when the bonus is paid. However, we feel that neither of these behavior patterns is likely to be widely observed. Notice that they violate either the ban on borrowing or the ban on dissaving. To the extent that these rules of thumb are used, imitating the behavior of S will be difficult. We

[9] However, a related issue is investigated by Landsberger (1966).

[10] This explains why many teachers sacrifice interest by electing to receive their academic-year salary paid in 12 monthly installments (September–August) rather than 10 (September–June).

believe that for the typical individual much of the bonus will end up being saved, especially through the purchase of durable goods. This amounts to the use of an auxiliary rule regarding the disposition of bonuses received. We expect that total saving for B will exceed that for S because of the technology of self-control. By paying the bonus the firm is acting as an external self-control device (much like parents who tell children that money gifts at Christmas must go into a savings account). Temptations to spend during the year will be overcome because the smaller monthly salary will make them seem beyond the individual's means. This technology seems to have been recognized by the millions of taxpayers each year who claim too few exemptions in order to assure a tax refund.[11] Obviously some self-deception (doer deception) is necessary for this device to work, but doers can apparently be deceived quite easily. How else can one explain the not uncommon practice of knowingly setting one's watch a few minutes ahead ("in order to get to places on time")?

As the analysis in this section suggests, the shape of the income stream will affect the type of saving strategy adopted. Those individuals with variable and uncertain incomes will find a discretionary rule such as saving s percent each month difficult to enforce in the low-income months. Without a mandatory saving plan, they would have to adopt some more complex strategy to save effectively. Similarly, those individuals whose incomes are expected to decline (such as professional athletes) would prefer to save a large proportion of their high current incomes and a smaller (perhaps negative) proportion of their lower future incomes. For both, a mandatory pension plan is particularly likely to increase total saving. A more sophisticated analysis of the effect of pensions on saving might detect differences of this sort.

The case of athletes points up the extreme differences in behavior that self-control can produce. Their declining income stream creates a difficult self-control problem in the high-income years. Some athletes hire agents to invest their incomes and limit their current spending, and many of them become rich. Others rely on discretionary strategies and end up bankrupt. Both types of behavior are possible in our model, depending on the degree of planner control and the types of precommitment strategies available. The best predictors of which individuals will fall into which groups are probably related to family background, since the family is the most likely place for the individual to learn (or not learn) the rules and norms necessary to overcome the self-control problems.[12]

[11] E.g., in 1969, 55 million taxpayers received refunds while 18 million owed taxes. In dollars, overpayments exceeded underpayments by 39 percent.

[12] The last two paragraphs were prompted by a suggestion from Sam Peltzman. Irving Fisher also discussed some of these issues. He notes that some individuals spend

C. *Individual Marginal Rates of Time Preference*

The orthodox theory of intertemporal choice as formulated by Irving Fisher produces a very strong result. Each person should equate his marginal rate of time preference (MRTP) with the relevant after-tax interest rate.[13] Thus the theory predicts that all individuals who face the same after-tax interest rate will make the same marginal intertemporal choices: Specifically, they will act as if they used the after-tax interest rate as their discount rate. This follows because individuals are assumed to use capital markets to arbitrage away any difference between what Fisher called their "rate of impatience" and the interest rate. It should be noted, however, that this result can fail to hold if capital markets impose quantity constraints on borrowers (capital rationing). In this case borrowers may be forced to stop borrowing even though their rate of impatience exceeds the interest rate.

In exactly the same fashion, self-imposed borrowing constraints such as those discussed in Section IIB prevent the complete internal arbitrage from taking place. Thus in our model, in which such constraints play an important role, the presumption that individual MRTPs will equal the interest rate no longer holds. Indeed, we expect to observe behavior that implies an MRTP greater than the interest rate and at the same time an unwillingness to engage in additional borrowing. Two points need to be raised about this implication. First, once rules are incorporated into the analysis, the failure to equate the MRTP to the interest rate may not violate an optimality condition. If rules are used it is because they lead to higher levels of utility than pure discretion would. The inequality created is costly, but the cost arises from the necessity of using a second-best technology. Because rules by nature must be simple, rules that select the precisely correct consumption bundle in every situation are infeasible. Second, notice that if a quantity constraint is binding, whether internally or externally imposed, observed MRTPs will be equal to or greater than the interest rate.

Attempts to measure individual MRTPs appear in studies by Kurz, Spiegelman, and West (1973) and Hausman (1979). Hausman studied families' purchases of room air conditioners. The trade-off between initial outlay and operating costs permitted him to estimate implicit MRTPs. The mean MRTP in his sample was about 25 percent, clearly above any relevant interest rate. Kurz et al. obtained similar results

---

their weekly paycheck at the "grog house." (Others, we believe, avoid the grog house precisely on those days.) He attributes much of the observed differences in behavior to social class. Our differences with Fisher are discussed in the next section.

[13] We define the MRTP to be the marginal rate of substitution between tomorrow's consumption and today's consumption minus one.

by asking hypothetical questions of participants in the Seattle and Denver Income Maintenance Experiments. They asked a sample of participants a series of questions of the following sort: What size bonus would you demand today rather than collect a bonus of $100 in 1 year? Several different forms of this type of question were asked, and the results were striking. For whites the mean rate of time preference implied by their answers varied between 36 and 76 percent. For blacks the rates varied between 40 and 122 percent. Of particular interest for present purposes is the fact that this sample included only those respondents who said they could borrow either $500 to make an installment purchase or $1,000 in cash. Furthermore, 81.3 percent of this subsample reported that they would not borrow $1,000 at current interest rates. (The mean perceived rate was generally less than 20 percent.) This strongly suggests the use of self-imposed borrowing constraints.

Once individual differences in MRTPs are anticipated, it becomes interesting to ask what factors determine those differences. A detailed examination is beyond the scope of this paper, but the model predicts that those factors determining individual rates of impatience will (for the reasons stated above) also affect observed MRTPs. These factors are discussed at length by Fisher (who draws on the writings of John Rae and Eugen Böhm-Bawerk). Fisher believed that age, income, and marital status affect the rate of impatience (in obvious directions). He felt that the shape of the income stream as well as the level was important; if income were expected to rise, the rate of impatience would be higher. Six "personal characteristics" were also deemed to be important: foresight, self-control, habit, expectation of life, concern for the lives of other persons, and fashion.[14]

In principle, the planner-doer model can be tested against the standard model quite simply. Observe some intertemporal choice that implies a discount rate (such as Hausman's study of air conditioners). Then regress the implied discount rate on the factors above *and* the individual's borrowing rate. The standard framework implies that only the borrowing rate will be a significant predictor. Our model implies that the other factors will also be important. Hausman did test one such variable. The implied rate of discount was computed for six income classes. His results appear in table 1. Clearly the variation in discount rates cannot be attributed solely to variations in borrowing rates.

We would also expect age and social class to be important in predicting individual intertemporal choices. The young behave impatiently in part because they have yet to master the techniques of

[14] Fisher 1930, p. 81. On this topic also see Maital and Maital (1977).

TABLE 1

ESTIMATED DISCOUNT RATES USING MEAN POPULATION ESTIMATES

Income Class ($/yr)	Observations (*N*)	Implied Discount Rate (%)
600	6	89
10,000	15	39
15,000	16	27
25,000	17	17
35,000	8	8.9
50,000	3	5.1

SOURCE.—Hausman 1979

self-control. To the extent that these techniques are learned from parents, class differences will be observed. Most important, our model stresses the theoretical admissability of these variables. Only further empirical work can establish their relative explanatory power.

IV. Summary and Conclusion

We now briefly recapitulate our argument. We have investigated intertemporal choice as a problem in the economic theory of self-control. As the quotation at the beginning of Section I states, the concept of self-control is paradoxical unless some kind of multiself model of man is adopted. We have introduced self-control into a formal model of intertemporal choice by modeling man as an organization with a planner and many doers. Conflict occurs because the doers are myopic (i.e., selfish). This conflict is fundamentally similar to the agency relationship between the employer and the employee, and individuals use many of the same strategies that organizations adopt to deal with their "conflicts of interest." These strategies can involve doer/employee discretion while their incentives have somehow been altered, or they may entail the implementation of precommitment (a rule) to avoid the doer/employee decision process altogether.

The close correspondence between the solutions to control problems adopted by organizations and individuals provides strong support for our model. Although our model is nontraditional, our tools are strictly traditional. Formally, our model closely resembles that used by Ross (1973) in his study of the theory of agency. Finally, we note that ours is a theory of rational behavior, just as Ross's theory is of profit-maximizing behavior.

Many applications of the model are possible, and we have discussed a few briefly. The most important applications are in the study of

individual saving behavior. Our hypotheses are quite different in spirit from the permanent income and life-cycle hypotheses that currently dominate the literature. On the basis of the evidence presented here, we feel these theories of saving should be reevaluated.

References

Ainslee, George. "Specious Reward: A Behavioral Theory of Impulsiveness and Impulse Control." *Psychological Bull.* 82 (July 1975): 463–96.

Berlin, Isaiah. "Two Concepts of Liberty." *Four Essays on Liberty.* New York: Oxford Univ. Press, 1969.

Blackorby, Charles; Nissen, David; Primont, Daniel; and Russell, R. Robert. "Consistent Intertemporal Decision Making." *Rev. Econ. Studies* 40 (April 1973): 239–48.

Buchanan, James. "The Samaritan's Dilemma." In *Altruism, Morality, and Economic Theory*, edited by Edmund S. Phelps. New York: Russell Sage, 1975.

Cagan, Philip. *The Effect of Pension Plans on Aggregate Saving: Evidence from a Sample Survey.* New York: Columbia Univ. Press (for Nat. Bur. Econ. Res.), 1965.

Elster, Jon. "Ulysses and the Sirens: A Theory of Imperfect Rationality." *Soc. Sci. Information* 16 (October 1977): 469–526.

Feldstein, Martin. "Do Private Pensions Increase National Saving?" Working Paper no. 186, Nat. Bur. Econ. Res., Cambridge, Mass., 1977.

Fisher, Irving. *The Theory of Interest.* London: Macmillan, 1930.

Freud, Sigmund. "Beyond the Pleasure Principle." In *The Standard Edition of the Complete Psychological Works of Sigmund Freud*, edited by James Strachey and Anna Freud. London: Hogarth, 1958.

Hammond, Peter J. "Changing Tastes and Coherent Dynamic Choice." *Rev. Econ. Studies* 43 (February 1976): 159–73.

Hausman, Jerry A. "Individual Discount Rates and the Purchase and Utilization of Energy-using Durables." *Bell J. Econ.* 10 (Spring 1979): 33–54.

Jensen, Michael C., and Meckling, William H. "Theory of the Firm: Managerial Behavior, Agency Costs, and Ownership Structure." *J. Financial Econ.* 3 (October 1976): 305–60.

Katona, George. *Private Pensions and Individual Saving.* Ann Arbor: Univ. Michigan Inst. Soc. Res., 1965.

Kurz, M.; Spiegelman, R.; and West, R. "The Experimental Horizon and the Rate of Time Preference for the Seattle and Denver Income Maintenance Experiments: A Preliminary Study." Menlo Park, Calif.: SRI Internat. Res. Memorandum no. 21, November, 1973.

Landsberger, Michael. "Windfall Income and Consumption: Comment." *A.E.R.* 56 (June 1966): 534–40.

McIntosh, Donald. *The Foundations of Human Society.* Chicago: Univ. Chicago Press, 1969.

Maital, Schlomo, and Maital, Sharona. "Time Preference, Delay of Gratification and the Intergenerational Transmission of Economic Inequality: A Behavioral Theory of Income Distribution." In *Essays in Labor Market Analysis: In Memory of Yochanan Peter Comay*, edited by Orley C Ashenfelter and Wallace E. Oates. New York: Wiley, 1977.

Munnell, Alicia H. *The Effect of Social Security on Personal Saving.* Cambridge, Mass.: Ballinger, 1974.

———. "Private Pensions and Saving: New Evidence." *J.P.E.* 84, no. 5 (October 1976): 1013–32.

Myers, Stewart C. "Determinants of Corporate Borrowing." *J. Financial Econ.* 5 (November 1977): 147–75.

Peleg, Bezalel, and Yaari, Menahem E. "On the Existence of a Consistent Course of Action When Tastes Are Changing." *Rev. Econ. Studies* 40 (July 1973): 391–401.

Pollak, Robert A. "Consistent Planning." *Rev. Econ. Studies* 35 (April 1968): 201–8.

Ross, Stephen A. "The Economic Theory of Agency: The Principal's Problem." *A.E.R. Papers and Proc.* 63 (May 1973): 134–39.

Schelling, Thomas C. *The Strategy of Conflict.* Cambridge, Mass.: Harvard Univ. Press, 1960.

———. "Egonomics, or the Art of Self-Management." *A.E.R. Papers and Proc.* 68 (May 1978): 290–94.

Scitovsky, Tibor. *The Joyless Economy: An Inquiry into Human Satisfaction and Consumer Dissatisfaction.* New York: Oxford Univ. Press, 1976.

Shefrin, H. M., and Thaler, Richard H. "Rules and Discretion in a Two-Self Model of Intertemporal Choice." Mimeographed. Ithaca, N.Y.: Cornell Univ., April 1980.

Smith, Adam. *Theory of Moral Sentiments.* London: Millar, 1759.

Stigler, George J. *The Theory of Price.* 3d ed. New York: Macmillan, 1966.

Strotz, Robert H. "Myopia and Inconsistency in Dynamic Utility Maximization." *Rev. Econ. Studies* 23, no. 3 (1955–56): 165–80.

Thaler, Richard H. "Toward a Positive Theory of Consumer Choice." *J. Econ. Behavior and Org.* 1 (1980): pp. 39–60.

Yaari, Menahem E. "How to Eat an Appetite-arousing Cake." Research Memorandum no 26, Center Res. Math. Econ. and Game Theory, Hebrew Univ., June 1977.

Comments

A Note on Loss of Control and the Optimum Size of the Firm

Antonio Camacho and William D. White

University of Illinois at Chicago Circle

This note has been inspired by a recent paper in this journal by Calvo and Wellisz (1978). Its main purpose is to suggest that, in constructing models to study the size of firms, more attention needs to be given to the conditions which bring about their existence.

In Section II of their article, Calvo and Wellisz prove a proposition that given certain assumptions about supervision, if a firm of any size exists, then loss of control will not impose finite limits on its size. Although their result is correct, they, as well as some other researchers working in this field, fail to give explicit reasons for firms to exist. However, as we will argue below, there are not convincing reasons, under their assumptions, for firms of any size to exist, thus making their case for the appearance of unbounded firms rather weak.

This result prompts us to suggest, as stated above, that more attention must be given in this area of research to the very conditions which bring about the existence of firms. For instance, models which, in the spirit of Williamson's (1967) paper, explicitly consider problems with information and communication, or models where it is shown that the activities of workers or agents may be coordinated more efficiently through extra market arrangements, may both produce realistic and interesting results regarding the size of firms.

We proceed as follows. First, we prove that, under Calvo and Wellisz's assumptions, firms will never exist if workers can be self-employed. Second, we show by means of a "setup" cost example that,

We are grateful to Joseph Persky for his comments on an earlier version of this note. Any errors which remain are, of course, our own.

[*Journal of Political Economy*, 1981, vol. 89, no. 2]
© 1981 by The University of Chicago 0022-3808/81/8902-0002$01.50

even if conditions which limit the opportunities of workers for self-employment exist, Calvo and Wellisz's assumptions regarding skills, utility functions, and productivity of workers seem to imply that there are more efficient arrangements to carry out production activities than the formation of firms. The note concludes with an intriguing conjecture and a suggestion.

I. No Limitations on Opportunities for Self-Employment

We will proceed now to prove that, if there are no limitations on workers' opportunities for self-employment, then under Calvo and Wellisz's assumptions firms will never exist. We will obtain this result by showing that, under their assumptions, a worker can always achieve at least as high a level of utility by working for himself as he can by working as an employee for any firm which makes a positive profit out of hiring him.

In their model, Calvo and Wellisz assume that the (Von Neumann-Morgenstern) utility index, U, of every worker is the same and depends on consumption, c, and effort, e; more specifically they assume:

$$U = u(c) - v(e), c \geq 0, 0 \leq e \leq 1, \tag{1}$$

where $u' \geq 0, v' \geq 0, u'' \leq 0$, and $v'' \leq 0$. (The last two conditions rule out risk-loving behavior.) Effort is assumed to be at a maximum when e equals one, while the worker is assumed to be completely idle when e is equal to zero.

Workers employed in firms are subject to monitoring where P is the probability that a worker will be detected shirking. Workers who are not detected shirking are paid a wage of w. Workers who are found shirking are paid we.

For any P and w, a worker will select that level of effort which maximizes his expected utility z:

$$z = P[u(we) - v(e)] + (1 - P)[u(w) - v(e)]. \tag{2}$$

Let $\bar{e}$ be the value of e which maximizes z.

Calvo and Wellisz assume also that the productivity of every worker is the same and equal to βe $(0 \leq e \leq 1)$ when working at the level of effort e. Given this assumption, a firm's expected revenue from employing a worker at a wage w and with a level of supervision P will be $\beta\bar{e}$. The expected cost to the firm of employing the worker will be the expected payments to the worker, $P(w\bar{e}) + (1 - P)w$, plus the expected cost of providing a level of supervision P. Thus, the total expected cost of employing a worker will be greater than or equal to $P(w\bar{e}) + (1 - P)w$. A necessary condition for the firm to make a positive expected profit is then

$$\beta\bar{e} > P(w\bar{e}) + (1 - P)w$$

or

$$w < \frac{\beta\bar{e}}{P(\bar{e} - 1) + 1}. \tag{3}$$

Let $E(\hat{U})$ be the maximum expected utility which a worker can obtain by working for a firm at a wage rate w and under a level of supervision P:

$$E(\hat{U}) = P[u(w\bar{e}) - v(\bar{e})] + (1 - P)[u(w) - v(\bar{e})]. \tag{4}$$

The expected utility of a worker when self-employed and applying himself to his work at the same level of effort $\bar{e}$ as in (4) is

$$E(U) = u(\beta\bar{e}) - v(\bar{e}). \tag{5}$$

Taking into consideration condition (1) and inequality (3), we can easily obtain

$$E(\hat{U}) \leq u(\beta\bar{e}) - v(\bar{e}),$$

which in view of (5) proves our assertion.

II. Limitations on Opportunities for Self-Employment

We will now discuss conditions that may limit workers' opportunities to be self-employed. To facilitate this discussion we will first enumerate the main assumptions of Calvo and Wellisz's model: Everyone has the same skills; everyone has the same utility function and, therefore, the same attitude toward risk; there is no specialization of labor, no need for communication of instructions or for coordination of activities; there are constant returns to scale when everyone applies the same level of effort to production activities.

Given these assumptions and the additional assumption that there are no institutional restrictions on self-employment, it appears that the only condition which might limit opportunities for self-employment for part of the population is the existence of some factor of production, say land, whose ownership is unevenly distributed. This situation might allow owners to pay workers (those who lack any endowment of this factor) a wage below their marginal product. But let us analyze this situation in more detail. Consider two cases: (i) only one individual owns all the land; (ii) more than one individual owns land.

In case i, an individual who is a monopolist may be able to start a firm and make total profits greater than β by hiring workers at a wage below their marginal product. Then, as Calvo and Wellisz rightly prove, the owner can increase his profits by adding layers to the hierarchical structure of his firm until he employs everybody.

But this is a rather cumbersome arrangement. Suppose that the total amount of land is L, the total number of workers is k, and the maximum profit the monopolist can extract by employing k workers is $\pi(k)$. By renting each worker a plot of land L/k at a price of $\pi(k)/k$, the monopolist can obtain the same amount of money as he would by establishing a firm, while workers will achieve a level of utility greater than or at least equal to the level they would obtain by working for the monopolist. Under these circumstances, firms are unlikely to be established.

Similarly, firms are not likely to be established in case ii either. Any owner who might start a firm always has the simpler option available, like our monopolist in case i, of renting his land to the workers of his hypothetical firm and charging them a rent equal to the maximum possible profit he could obtain from establishing a firm divided by the number of his employees. Under this arrangement, the owner will receive the same amount of money as he would receive if he set up a firm, while workers will obtain a level of utility which is greater than or at least equal to the level of utility they would obtain by working for the owner as employees.

The results obtained by Calvo and Wellisz in section II of their 1978 paper and our own discussion above lead us to make the following conjecture, which we present as a still unproved theorem: that for any set of conditions under which there is no limit to the size of firms, there is no reason for firms to exist. And this conjecture, together with the fact that firms do exist and are of limited size, in turn suggests to us that the optimal size of firms may be determined by those very conditions that bring about their existence.

References

Calvo, Guillermo A., and Wellisz, Stanislaw. "Supervision, Loss of Control, and the Optimum Size of the Firm." *J.P.E.* 86, no. 5 (October 1978): 943–52.

Williamson, Oliver E. "Hierarchical Control and Optimum Firm Size." *J.P.E.* 75, no. 2 (April 1967): 123–38.

Prior Information and the Observational Equivalence Problem

Warren E. Weber

Virginia Polytechnic Institute and State University

The observational equivalence problem pointed out by Sargent (1976) consists of two parts. The first is whether regression analysis of so-called classical and Keynesian reduced-form equations can discriminate between the competing hypotheses, using data from a single policy regime in the absence of any prior information concerning the policy rule being followed. The answer to this question is no and can be shown as follows.

Consider the reduced-form equation for a "goal" variable (e.g., real GNP). The classical version of this reduced-form equation would be

$$\alpha(L)y_t + \beta(L)(m_t - E_{t-1}m_t) = \mu_t, \tag{1}$$

where m_t is a policy variable (e.g., the nominal money supply), and μ_t is a white-noise error term. This equation is classical, since only innovations in the policy will affect the goal variable. As a consequence, the time path of the goal variable will be independent of the policy rule chosen, so long as the rule implies $m_t = E_{t-1}m_t + \epsilon_t$, where ϵ_t is a white-noise error distributed independently of μ_t.

The Keynesian version of this reduced-form equation would be

$$a(L)y_t + b(L)m_t = \nu_t. \tag{2}$$

Assuming that (2) is invariant with respect to interventions that change the m_t process, it is broadly "Keynesian," since it would be possible to solve (2) for a policy feedback rule which would minimize the deviations of the goal variable around some desired level.

Professor of economics, Virginia Polytechnic Institute and State University. I am indebted to Anatoli Kuprianov, Bennett McCallum, and an anonymous referee for useful comments on this paper.

[*Journal of Political Economy*, 1981, vol 89, no 2]
© 1981 by The University of Chicago. 0022-3808/81/8902-0010$01.50

To show that estimation of (1) and (2) will not distinguish between the Keynesian and classical positions, suppose that the policy variable follows the rule

$$\gamma(L)m_t = \epsilon_t \tag{3}$$

over the entire observation period. Then substituting (3) into (1), we obtain

$$\alpha(L)y_t + \beta(L)\gamma(L)m_t = \mu_t. \tag{4}$$

In the absence of any prior information on the order of the polynomials $\gamma(L)$ or $\beta(L)$, it would be impossible to distinguish empirically between (4) and (2). That is, if both equations included the same number of lagged terms, they would have the same sum of squared residuals, since $\beta(L)\gamma(L)$ would be an arbitrary polynomial in L. Further, since (4) has an identical error to (1), the sum of squared residuals obtained from estimating (1) would be the same as that from estimating (2), regardless of whether the Keynesian or classical hypothesis were true. Thus, in the absence of any prior information on the lengths of the distributed lags in the model, it is impossible to empirically distinguish between the Keynesian and classical positions using data from a single policy regime.

The second question is whether such a test can be performed when prior information on the order of $\gamma(L)$ or $\beta(L)$ is available. McCallum (1979) has pointed out one such case. Our purpose is to generalize McCallum's result and to relate it to Sargent's original paper and the empirical tests of the classical and Keynesian hypotheses performed by Barro (1977, 1978).

Let $c(L)$ be the polynomial in the lag operator L of the regression coefficients on the m_t terms when equation (4) is estimated; that is, $c(L) = \beta(L)\gamma(L)$. Thus, (4) can be empirically distinguished from (2), if and only if there exist testable restrictions on the relationship between $\gamma(L)$ and $c(L)$. To examine the nature of these restrictions, first note that a typical element of the convolution $c(L) = \beta(L)\gamma(L)$ is

$$c_j = \sum_{k=-\infty}^{\infty} \beta_k \gamma_{j-k}.$$

Thus, $c(L) = \beta(L)\gamma(L)$ can be written in matrix form as

$$C = H(\gamma)B, \tag{5}$$

where $H(\gamma)$ is the $(n + m + 1) \times (n + 1)$ matrix

$$H(\gamma) = \begin{bmatrix} \gamma_0 & 0 & 0 & \cdots & 0 & 0 \\ \gamma_1 & \gamma_0 & 0 & \cdots & 0 & 0 \\ \gamma_2 & \gamma_1 & \gamma_0 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \gamma_n & \gamma_{n-1} & \cdot & \cdots & \gamma_1 & \gamma_0 \\ \gamma_{n+1} & \gamma_n & \cdot & \cdots & \gamma_2 & \gamma_1 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 0 & \gamma_m \end{bmatrix}, \tag{6}$$

$B = (\beta_0, \beta_1, \ldots, \beta_n)'$ is the vector of coefficients of $\beta(L)$, and $C = (c_0, c_1, \ldots, c_{n+m})'$ is the vector of coefficients of $c(L)$. Then by the result of Rothenberg (1973, p. 37), knowledge of m or n will yield restrictions relating the coefficients of $\gamma(L)$ and $c(L)$ if rank $[H(\gamma)] < (n + m + 1)$.

With this result, we can now interpret Sargent's original argument and McCallum's result. Sargent considers the case in which $n = m = \infty$. In this case, the restriction that the rank of $H(\gamma)$ be less than $n + m + 1$ is not satisfied, and no restrictions between $\gamma(L)$ and $c(L)$ are obtained. McCallum considers the case in which $n = 0$, $m = \infty$. In this case rank $[H(\gamma)] = 1 < \infty$, and the prior information given by knowledge of $\gamma(L)$ implies restrictions on $c(L)$.

Rothenberg's result, however, allows a more general statement of when knowledge of n or m implies that restrictions between $\gamma(L)$ and $c(L)$ exist. This general result is that knowledge of n or m will imply that testable restrictions between $\gamma(L)$ and $c(L)$ exist, unless $n = \infty$ or n is finite and $m = 0$. This generalization follows directly from the rank condition stated above. It implies that unless y_t depends on all past innovations in policy or policy is white noise, the classical reduced-form equation (4) can be tested against the Keynesian reduced-form equation (2). The test is whether imposing the restrictions of equation (5) leads to a statistically significant decrease in the explanatory power of the regression equation.

The test of the classical model by Barro used the policy equation

$$\gamma(L)m_t + \lambda(L)z_t = \epsilon_t \tag{3'}$$

in place of (3), where z_t is an exogenous variable (such as the minimum wage or level of unemployment compensation) and $\lambda(L)$ is a polynomial of order r in L with $\lambda_0 = 0$. Substituting (3') into (1) would yield the new classical reduced form

$$\alpha(L)y_t + \beta(L)\gamma(L)m_t + \beta(L)\lambda(L)z_t = \mu_t \tag{4'}$$

to be compared with the appropriately modified Keynesian version

$$a(L)y_t + b(L)m_t + d(L)z_t = \nu_t. \tag{2'}$$

Now let $f(L)$ be the polynomial in L of the regression coefficients on the z_t term when (4′) is estimated, so that $f(L) = \beta(L)\lambda(L)$. Then using the same arguments as above, (4′) can be empirically distinguished from (2′) if there exist testable restrictions between $\gamma(L)$ and $c(L)$ or between $\lambda(L)$ and $f(L)$. The case of $c(L)$ was considered before. Following the reasoning used for that case, we find that knowledge of r will yield testable restrictions between $\lambda(L)$ and $f(L)$ if and only if rank $[H(\lambda)] < (n + r + 1)$. Since $H(\lambda)$ is $(n + r + 1) \times (n + 1)$, knowledge of r will not imply that testable restrictions between $\lambda(L)$ and $f(L)$ exist when $n = \infty$. Thus, Barro's method of including exogenous variables in the policy rule which do not appear in (1) will not yield testable restrictions to distinguish classical and Keynesian reduced forms when y_t depends on all past innovations in the policy. However, when n is finite and $m = 0$, so that no testable restrictions between $\gamma(L)$ and $c(L)$ exist, the knowledge that $r \geqslant 1$ implies that testable restrictions between $\lambda(L)$ and $f(L)$ exist. Thus, in this case, Barro's method will permit the classical and Keynesian reduced-form equations to be distinguished when they could not be if z_t were omitted from the policy rule.

The practical usefulness of these results, of course, depends on the existence of reliable theoretical restrictions on the order of the lag distributions of $\beta(L)$, $\gamma(L)$, or $\lambda(L)$. The order of $\beta(L)$ depends on the extent to which past errors in the expectation of policy affect the current level of the goal variable. One source of restrictions on the value of n can be found in the models of aggregate supply developed by Lucas (1972, 1973) and Leiderman (1979) in which producers in informationally separated markets face the problem of distinguishing relative price changes from absolute price changes, given current information only on the prices they receive. In these models, output is affected only by the current error in the expectations of the price level implying $n = 0$. The same restriction on the value of n arises from Sargent's (1978; 1979, chap. 16) models of employment. McCallum also argues in favor of such a restriction on the grounds that "it is hard to imagine ways in which past expectational errors could have direct effects on current behavior—bygones are, after all, bygones" (1979, p. 398). Thus, there appears to be a solid theoretical basis for imposing the restriction that $n = 0$ on $\beta(L)$.[1]

[1] This conclusion is not altered by the consideration of adjustment or search costs. Such costs may lead to past expectational errors having indirect effects on the current

Given $n = 0$, the rank conditions above indicate that testable restrictions between $c(L)$ and $\gamma(L)$ or between $f(L)$ and $\lambda(L)$ will exist unless policy is white noise (i.e., unless $m = r = 0$). As yet, there is no satisfactory theory of how policy functions are determined, which would allow us to theoretically determine that policy is not white noise. However, the evidence from empirical attempts to estimate monetary policy functions does support such a conclusion.[2]

Thus, we would argue that reliable theoretical and empirical evidence exists to allow testing of models with classical implications against models with Keynesian implications, using data from a single policy regime. However, the lack of theoretical restrictions on m or r does caution that such tests should check for their robustness to changes in the length of the lags assumed in the policy functions underlying the test.

References

Barro, Robert J. "Unanticipated Money Growth and Unemployment in the United States." *A.E.R.* 67 (March 1977): 101–15.

———. "Unanticipated Money, Output, and the Price Level in the United States." *J.P.E.* 86, no. 4 (August 1978): 549–80.

Froyen, Richard T. "A Test of the Endogeneity of Monetary Policy." *J. Econometrics* 2 (July 1974): 175–88.

Leiderman, Leonardo. "Expectations and Output-Inflation Tradeoffs in a Fixed-Exchange-Rate Economy." *J.P.E.* 87, no. 6 (December 1979): 1285–1306.

Lucas, Robert E., Jr. "Expectations and the Neutrality of Money." *J. Econ. Theory* 4 (April 1972): 103–24.

———. "Some International Evidence on Output-Inflation Tradeoffs." *A.E.R.* 63 (June 1973): 326–34.

McCallum, Bennett T. "On the Observational Inequivalence of Classical and Keynesian Models." *J.P.E.* 87, no. 2 (April 1979): 395–402.

Neftci, Salih, and Sargent, Thomas J. "A Little Bit of Evidence on the Natural Rate Hypothesis from the U.S." *J. Monetary Econ.* 4 (April 1978): 315–19.

Rothenberg, Thomas J. *Efficient Estimation with A Priori Information.* New Haven, Conn.: Yale Univ. Press, 1973.

Sargent, Thomas J. "The Observational Equivalence of Natural and Unnatural Rate Theories of Macroeconomics." *J.P.E.* 84, no. 3 (June 1976): 631–40.

———. "Estimation of Dynamic Labor Demand Schedules under Rational Expectations." *J.P.E.* 86, no. 6 (December 1978): 1009–44.

———. *Macroeconomic Theory.* New York: Academic Press, 1979.

value of the goal variable, because they would affect its value in previous periods. However, such effects would be captured in $\alpha(L)$ (see McCallum 1979, p. 398).

[2] See Froyen (1974), Neftci and Sargent (1978), and the references contained therein for estimates of monetary policy functions which support this conclusion.

Book Reviews

The Enclosure of Ocean Resources: Economics and the Law of the Sea. By Ross D. Eckert.
Stanford, Calif.: Hoover Institution Press, 1979. Pp. 390. $16.95.

This is a clearly written book about the Law of the Sea, applying Coase/Demsetz economic ideas to regulation of, and property in, six or more uses of ocean resources The central theme is that national "enclosure" by the extension of base lines, and widening of territorial seas and various types of exclusive economic zones, are analogous to the historic "enclosure" of common lands and the appropriation of frequencies and so on. The book closes with chapters on UNCLOS (the United Nations Conference on the Law of the Sea) as a venue in which enclosure is debated and arranged.

Written for a more general audience than academic economists, the book is sometimes less than rigorous in its analysis of rents, profits, and transactions costs and may also be given to sliding over some difficult details of law, geography, engineering, and the other fields in which expertise is needed to write such a book satisfactorily to all. But the documented attempt to put the claims of an economics/law property rights approach is worth pondering.

The author fully realizes that "watertight" exclusivity arrangements for some ocean uses would be astronomically costly or even technically impossible. Consider fisheries, to which he devotes a long chapter. Enclosure of coastal fisheries leads to internalized control over local stocks but is obviously less reliable in controlling the harvesting and growth of migratory species. For the latter, the author faintly recommends (as he does for some types of oil-spill problems) a mixed international-coastal regime. But he warns that, for such arrangements, the transactions costs between nations are high; he conjectures that costs for cooperation, enforcement, and decision making increase seriously with the number of countries participating; and he comes out in favor of single-coastal-nation fisheries' regimes.

In this recommendation he is in distributive deep water, as he recognizes. The alternative to single-nation control is treaty-organized "communal" control, giving some power to all nations. Eckert argues that this system is inferior to that by management by individual states. The evidence of inferiority is only one of cost and effectiveness: that transactions costs do seem high, and success does seem incomplete, in such bodies. On such cost evidence, the author seems inclined simply to argue that single-nation regimes should prevail. But this alternative is not one that is available. Not only are many ocean resource uses unenclosable physically, but many claims to conflicting fishing or other rights are of ancient vintage and cannot be cancelled merely by an economist

[*Journal of Political Economy*, 1981, vol. 89, no. 2]
© 1981 by The University of Chicago

asserting that fishing success is higher and costs lower when one country takes all.

There *is* an answer, involving multinational equity ownership of the rights to use certain ocean resources, actual management being entrusted by all to one nation or to a private firm. The Pacific fur seal treaty provides an example. Some UNCLOS proposals having to do with minerals and other resources could be interpreted or remolded in this form, but Eckert does not do so. His adamant opposition to large numbers is therefore unconvincing. (A side benefit of this discussion is his analysis of the process of national adjustment to the likely future outcome of UNCLOS bargaining. A country about to make a bargain that depends upon its initial position tries to magnify its initial position. Eckert notes this several times.)

A second answer to the many-nation difficulty is to offer compensation to those who are excluded. If it is too costly to let all nations fish or mine, then Eckert should provide for some way to buy them out. Although the process of doing this would be costly, it might work, and the UNCLOS pursuit of the dogma of the oceans being "the common heritage of mankind" could be utilized here. But Eckert seems to me to give little if any weight to the view that UNCLOS is a constitution-making body from which all must depart better off than they arrived.

In other words, his quickness and clarity in making points against UNCLOS procedures do, it seems to me, blind him to the difficulty that distributional imperatives do call for multination ownership or benefit. Otherwise the injustice of exclusion would match that of the original enclosure movements, which Eckert seems to ignore.

In economists' jargon, Eckert seems strongest on efficiency and weakest in the realms where distribution and public choice meet. This is a pity, for his book ends in frustrated anger at the "hypocrisy" of bargaining at UNCLOS, when his earlier treatment, clean and informative, somewhat extended beyond the domain of operating transactions costs, would suggest the limits of what is possible given the costs of constitution making and the conflict in every nation's breast of distributional and efficiency objectives.

ANTHONY SCOTT

University of British Columbia

Issues in Financial Regulation. Edited by FRANKLIN R. EDWARDS. Regulation of American Business and Industry Series.
New York: McGraw-Hill Book Co., 1979. Pp. 526. $42.50.

This book can best be thought of as an intellectual sampling of approaches to regulatory problems by a diverse group of economists and lawyers. For those who want such a sampling, reading it would be useful. All economists do not think one way about regulation, and all lawyers another. There is a diversity of opinion in both camps. As an economist, I will venture the opinion that the economics profession can offer two criteria that should be applied to analyses of regulatory problems: (1) Are sound theoretical notions applied to the problems? and (2) Is good empirical methodology used? The first criterion involves not merely an understanding of theory but an understanding of the problem, how it arises, and the market forces that surround it. The second criterion involves not merely the gathering and analyzing of information but

asking the right questions about the type, amount, and quality of information needed to analyze the problem. It is useful to look at the papers in the book in this manner. Lawyers will undoubtedly have other criteria, but in my opinion a great number of our regulatory problems have arisen because one of these two criteria has been violated.

The papers in this book indicate that some thinkers pay attention to these criteria, others do not, and the quality of the papers is influenced accordingly. Kenneth Scott analyzes the dual banking system as a model of competition among regulatory agencies. By dual banking system, he means the system of alternative chartering at the state and national level and the system of divided regulation at the federal level that goes along with it. However, he only analyzes one-half of the problem. Competition among regulators lets banks prevent the imposition of unduly burdensome regulation by using the option of a change in chartering status to play one regulator off against another. Scott says this is good. But unless there are other checks on the regulators, the logical end of such a process is no regulation at all—that is, competition in laxity. But other checks do occur. Bank regulation and supervision is a technical business in which the best systems and procedures are far from obvious even to a trained professional. Regulators, like the rest of us, do not like failures, and they learn from each other's mistakes. As in the marketplace, alternatives in bureaucratic regulation promote innovation and efficiency, a form of "competition in excellence." In their periodic oversight hearings on the condition of the banking system, the Congressional banking committees are not just assessing the credibility of the reports of one monolithic regulatory agency. They are comparing the performance of several agencies. It is not obvious that the system produces overregulation or underregulation. But there are inherent checks and balances against either tendency.

Bank regulators have the dual responsibility of preventing extensive bank failures which would bring about unwanted contractions in the money supply and of insuring low-cost delivery of banking services to the public. This dual responsibility should be kept in mind when analyzing proposals for change. Scott presents an extremely complicated organizational chart to depict the structure of the system that regulates commercial banks. The chart displays the interrelationships between the Comptroller of the Currency, the Federal Reserve, the Federal Deposit Insurance Corporation, and the 50 state regulatory agencies. The chart is incomplete. It does not include the Securities and Exchange Commission, the Federal Trade Commission, the Treasury Department, the Department of Housing and Urban Development, the Commerce Department, the Labor Department, and the Energy Department—some of the other agencies that have gotten into the business of regulating banks. In addition, the Federal Home Loan Bank Board, the National Credit Union Administration, and the Farm Credit Administration—all omitted from the chart—are also regulators of deliverers of banking services and interact with the other bank regulators. The chart was constructed by Senator Proxmire's staff for the purpose of justifying consolidation and simplification of federal bank regulation. If we wish to design bank regulatory systems for the administrative convenience of the regulators, the senator's proposed reforms merit consideration. But if the purpose of the regulatory system is to insure continuous and efficient delivery of banking services in the marketplace, the failures of the current system are not obvious.

Failure to apply both criteria is illustrated in Roy Schotland's paper on conflicts of interest. He relies extensively on anecdotal evidence and con-

structed situations with very little analysis of market structure or economic incentives. On page 125 he says that most conflicts occur because competition is not effective, an assertion that is not justified with any evidence. On page 128 he hastens to add that most conflict situations are not exploited. Could it be that competition is effective? He says that confidentiality requirements make "marketplace type shopping" unfeasible for doctors, lawyers, and investment managers. No evidence is offered on the extent to which consumers actually do shop around for the services of such individuals. I make an attempt at it, and I do not think I am that unusual. Schotland makes an issue out of the conflict an investment banker incurs because of his obligation to "the buying public, to corporate issuers, and self-interest." He does not say why this is any different from the problem faced by the president of a manufacturing firm who has to mediate the competing claims of customers, suppliers, and stockholders, or why it should require any more regulation.

The paper by Robert Shay, "Consumer Protection in Credit Markets: An Analysis of Regulatory Reform," is an excellent example of judicious application of both criteria. Shay uses the economic analysis of property rights to assess the regulatory reform movement affecting consumer credit markets. He develops an analytical framework and applies it to one limitation of a creditor's remedy to recover debt: the prohibition of wage assignments. He then compares his analysis of the problem to that used by the Federal Trade Commission (FTC). The comparison is fascinating and revealing. Although he does not do any empirical testing, he develops a precise way of getting answers to the questions needed to implement his theory.

Shay believes a problem in consumer credit markets today is that the focus of bargaining is on rates and terms and not on the remedies and rights of creditors and borrowers. Rather than prohibit wage assignments, parties should be permitted to bargain over whether or not they are to be allowed in the contract. The author notes that when wage assignment provisions are allowed they are not uniformly found in all credit contracts. This suggests that when they are included, the reason is to lower the cost of credit to high-risk applicants.

To assess the need for government interference Shay suggests using Demsetz's principle of compensation. The government should try to buy permission to allow the activity from those who feel they would be harmed by allowing it, and it should try to buy permission to restrict the activity from those who feel they would benefit from allowing it. The lower-cost option should be purchased.

Shay's analysis indicates that the adversary relationship is not between debtor and creditor but between those who will be harmed if there is freedom to contract for remedies and those who will be harmed if there is not. The creditor is best viewed as an intermediary. Only a modest portion of the funds being lent are his own funds. Nevertheless, the creditors' adjustments to governmental intervention determine who gets helped and who gets hurt. The need is for marginal benefit/cost comparison between two groups of debtors, not between the debtors vis-à-vis the creditors. The rather crude balancing process used by the FTC is between the creditors and the debtors.

Two of the papers dealing with electronic funds transfer (EFT) propose radical policy changes without using either of the criteria mentioned above. Almarin Phillips claims that regulation and institutional arrangements—interest rate ceilings, deposit insurance, etc.—cause difficulties in the administration of monetary policy. Undoubtedly this is true. But he then goes on to assert, in language phrased to make the dissenter feel antediluvian, that EFT

makes all of this substantially worse and will bring a crisis. He offers no evidence or analysis to support his conclusion, and I, for one, dissent from it. To be sure, many regulations cause difficulties for private bankers, central bankers, and the public. But it is the regulations that cause the problem, not the use of electronic forms of communication.

Similarly, Alan Westin says that EFT creates the need for new statutes to spell out individuals' rights. He says that when information is merged from a variety of sources and not used for one-time purposes a new trustee relationship is created. Of course, such information gathering is as old as banking itself. But when it is done electronically, Westin believes new laws are called for. He is not exactly sure about what the new laws and regulations should be. He proposes a decade or two of controlled experimental regulation. He gives no discussion of current abuses to support this recommendation.

Arnold Heggestead does a survey of studies of bank competition and performance. Most of the articles surveyed, and the survey itself, suffer from a lack of application of the criteria mentioned here. This is best illustrated by Heggestead's ambiguous and contradictory conclusion. He starts out by stating that local banking markets are "more monopolistic than competitive." He ends by stating five major areas where further empirical research is needed to clear up anomalies.

If he had started out by analyzing the way markets operate he could have explained some of the anomalies himself. Interinstitutional competition is a fundamental fact of today's marketplace. If one examines all of the various asset powers, liability powers, and fee income powers of commercial banks, there is only one privilege that commercial banks have that is not granted to at least one other financial institution. This is the power to offer demand deposits, a privilege whose value is rapidly being eroded by market innovations and changing product mixes. In 1947, commercial banks held 56.3 percent of the assets of all financial institutions. In addition to thrift institutions, this latter category includes life insurance companies, private pension funds, state and local employee retirement funds, money market funds, and finance companies. In 1977 the corresponding figure was 39.6 percent. In 1947, they held 81.7 percent of the total deposits of all depository institutions. In 1977, it was 59.3 percent. In 1947 commercial bank deposits were 56.2 percent of the total liabilities of all financial institutions. In 1977 the corresponding figure was 32.4 percent. Demand deposits, the product for which commercial banks are sometimes said to have a "monopoly," were 40.3 percent of total financial institution liabilities in 1947. In 1977, the corresponding figure was 10.9 percent. This is not the picture of a monopolistic industry. When they had reason to do so, bank customers were highly sophisticated at finding alternative ways of meeting their financial needs.

Some of the other papers are good examples of applications of the criteria set forth here. Others are not. But enough has been said to illustrate the problem.

The book suffers from a lack of discussion of the political aspects of regulation. I refer to the process whereby information is gathered through hearings and testimony, and laws are passed; and more information is gathered through hearings and testimony, and regulations are written. With increasing frequency, outsiders are having a hard time matching the laws with the regulations, and Congress and the regulators are pointing their fingers at each other. I have no final answer to this problem, but I will suggest two partial causes. First, the process inherently favors anecdotal evidence over econometric work and surveys which, because they are subject to

methodological criticisms, inevitably present a more balanced view of the world. Second, when the banking system works well, the bankers get the credit, not the regulators. When it goes awry, the regulators share a good portion of the blame. This inevitably promotes regulations designed to protect inefficient banks, to the detriment of the customers of well-managed banks.

P. MICHAEL LAUB

American Bankers Association

New Challenges to the Role of Profit. Edited by BENJAMIN M. FRIEDMAN. Lexington, Mass.: D. C. Heath & Co., Lexington Books, 1978. Pp. viii+127. $14.95.

In 1976, Professors Samuelson, Arrow, and Lundberg presented lectures in the third series of the John Diebold Lectures. Along with an introductory chapter by Professor Friedman and a concluding round-table discussion, these lectures comprise *New Challenges to the Role of Profit.* Samuelson gives a broad-brush treatment of the trend of profits, and intellectual opinion about profits, in the modern mixed economy. He also includes a brief history of profit theory. Arrow links criticism of profits to general attacks on private property, but he emphasizes the conceptual differences between profit, property income, and private property. Lundberg reports on the Swedish "middle way"—now dubbed "fund socialism"—and attempts to draw conclusions for Western countries generally.

All things considered, the book is a disappointment from almost any perspective. Even for a work of this genre, *New Challenges* lacks focus. Though edited by a perspicacious scholar who contributes an introductory essay, the book wanders. The subject remains unclear. The principal lectures were delivered by academics, with comments by a mixture of academics and nonacademics. At a minimum, one would have hoped for careful delineation of a profit concept, distinguishing profit from equilibrium returns to owners of capital, land, and other scarce resources (i.e., interest and economic rent). Though any number of allusions and some specific references are made to the distinction between interest (or ordinary profit), rent, and pure profit, these various incomes are constantly conflated. The analysis is further confused by unwarranted shifting in the lectures between static and dynamic analysis. For instance, in any essay containing a number of fruitful insights, Friedman observes that "real economies are dynamic, not stationary" (p. 4); he then notes the importance of this realization for our understanding of entrepreneurship, innovation, development, and industrial structure. This discussion is immediately followed, however, by a reference to the intertemporal allocation role of profits (more properly an interest problem). Profits are then treated as accruing to owners of capital, thus entangling pure profits, rents, and interest. Shortly thereafter, Friedman considers the trade-off between equity and efficiency—or distribution and allocation—arising in a profit system. Almost every contributor eventually refers to this trade-off. The discussions of the "trade-off" remind one of those concerning the "cruel choice" between inflation and unemployment. There may indeed be different combinations of income distributions and allocation efficiency, depending on the taxes and other constraints placed on profits, but it is by no means clear that any stable social choice set exists from which one could pick and choose

combinations. Further, if one considers the behavior of economies over time, then the question is not whether profits (however defined) contribute to income inequality, but whether the profit *system* produces more or less income equality compared to other, *feasible* economic systems. The latter question is not really addressed in this volume.

Along with Arrow's, Friedman's is one of the better contributions to the book. The shifting from dynamic to static analysis and between profit concepts, even in the introductory chapter, is indicative of problems in the rest of the book. No fruitful analysis of "profit," or of any other concept, is possible without a shared understanding of the nature of the problem being discussed. Such a shared understanding is lacking in *New Challenges*. The contributors could surely have adopted a number of profit concepts or definitions. But regardless of the concept or definition chosen, reasonable consistency in analysis is necessary. In this book there is almost no analysis of profits as the essential signaling and motivating force in a market economy. Despite obligatory references to Schumpeter, no approach treating profit as an income emerging from unexpected changes in economic data is seriously considered or analyzed. The work of Kirzner, identifying genuine profit as a return to superior entrepreneurship—alertness to altered conditions—is totally ignored (Kirzner 1973, 1979). Yet if the related theories of Kirzner, Mises, Schumpeter, and Knight are at all correct, then profit—not some fictitious auctioneer—is responsible for any tendency toward plan coordination and equilibrium in decentralized economies. Coordination is, of course, always incomplete, because change is always occurring. But without appropriable profits, the decentralized market system cannot work.

The Kirznerian profit theory clarifies a number of conceptual problems. Profit, being a return to a special labor skill, is no longer conflated with interest, a return associated with capital accumulation. Nor is profit a rent or quasi rent accruing to a fixed or temporarily fixed factor, as in standard textbook analysis. Since profit disappears as quickly as it is publicly identified, and since it is not an economic rent, profit is not a taxable surplus. Arrow and Lundberg suggest that the ownership of property and the receipt of property income can be separated without serious allocational consequences. As Arrow put it: "This justification of the charging of price for the use of property is not necessarily an argument that the reward should go to the owner rather than to someone else" (p. 53). This line of argument is more plausible if profit is considered as a residual or rentlike income. In Kirzner's analysis, however, profits are the driving force of economic activity. They are actively and consciously strived for. They neither accrue to passive resource owners nor are they a residual. Thus, profit income and private property are inextricably linked and intertwined. As Marxists have always understood, the fundamental issue is the private property system and the very existence of markets.

The latter point helps indicate the deficiency of the book. Most of the challenges to profits discussed are quite old. Yet the contributions do not build on the long literature on the subject (Arrow's essay again is the exception). Newer critiques of profit and the market system derive from "highbrow opinion" and not from popular concerns (to which the book is ostensibly addressed). The ordinary person is not concerned that market economies do not attain global maxima and are never in general equilibrium. (Whoever suggested they would be?) The ordinary citizen is becoming aware, however, of the tendency for governmental policy in a mixed economy to substitute a system of monopoly rents and arbitrary incomes for that of earned profits (see Hughes 1977). He may lack the theoretical background to explain the

process, but the man in the street is surprisingly on the mark in his judgments. Lundberg tells us that Western economies will, like Sweden's, become more mixed (p. 91). If so, monopolization and socialization of earned income will become more widespread.

In the mixed economy, or fund socialism, the criticism that profits do not motivate innovation or lead to plan coordination changes from an erroneous critique to an essentially correct one. Inter alia, as measured profits are taxed at higher rates and supplanted by explicit and implicit subsidies, their motivational and allocational importance is diminished. Yet this process will remain largely hidden if viewed from a perspective in which profit and economic rent cannot be clearly distinguished in an analytical sense. In the latter view, "profit" is a type of surplus that, at least within some range, can be altered with little impact on resource allocation.

If, in fact, this book were to help focus attention on the distinctive operation of a mixed, as opposed to a market, economy, then it would have made a contribution. Were the book to help spark a public debate on this question, its faults could be overlooked. But if such a debate were to occur, it would probably be more in spite of than because of such efforts as this book.

GERALD P. O'DRISCOLL, JR.

New York University

References

Hughes, J. R. T. *The Governmental Habit: Economic Controls from Colonial Times to the Present.* New York: Basic, 1977.

Kirzner, Israel M. *Competition and Entrepreneurship.* Chicago: Univ. Chicago Press, 1973.

———. *Perception, Opportunity, and Profit: Studies in the Theory of Entrepreneurship.* Chicago: Univ. Chicago Press, 1979.

Economic Equality and Fertility in Developing Countries. By ROBERT REPETTO. Baltimore: Johns Hopkins University Press, 1979. Pp. 206. $14.50.

This work is best viewed as a collection of pieces in which a model of the relationship between income and fertility is developed, applied, and in turn tested through these applications. The questions raised by the author are crucial ones for population policy, his answers are well reasoned, and if there is still room for disagreement regarding his policy conclusions it is nevertheless much to his credit that the evidence and the issues are well laid out in the book itself.

The book opens with a statement of its central theme: that fertility is a nonlinear negative function of income such that the marginal effects of an income change are greater for poor people (although perhaps not for the abject poor) than for the rich. It follows from this nonlinearity that a given change in income level would have little impact on fertility for families in the middle and upper end of the income distribution but might very well have a large impact on fertility for poor families. The policy implication is that redistribution of income from rich to poor would generate a decline in aggregate fertility and hence population growth.

Having stated this hypothesis in chapter 1, chapter 2 develops a cross-country model for testing its implications. Here it is necessary to assume that the basic structural relationship between fertility and income at the family level is essentially the same in every country (*ceteris paribus*). If so, it follows quite easily that any redistribution of income from the relatively rich to the poor, including transfers from rich countries to poor countries, would reduce aggregate fertility. It also follows that the relationship between aggregate fertility and income is not fully specified by the level of national income without some measure of its distribution. Chapters 3–5 are country studies establishing the robustness of the basic relationship. Chapter 6 concludes the book with an analysis of worldwide fertility along the lines suggested above.

Each of the country studies uses some multivariate analysis based on household data as well as more aggregated descriptive statistics. In general, the regressions are less satisfying than the descriptive analyses. In part this may be because of difficulties obtaining data for the appropriate concepts and interpreting results based on imperfect data. Largely, however, it is because the separate country studies have not been well integrated into a coherent whole. Although the relationship between income and fertility is the only point at issue, a fairly complex country-specific model is needed in order to isolate that relationship. Despite the claim that chapter 2 develops a general model, each chapter contains its own model which is not clearly related to the "general" one. Moreover, the development of these models, the discussions of the data, and the interpretation of results are all too cursory to be satisfying, even though the limited purpose of this study may not warrant a full country analysis for each case. The author's analysis is much more convincing when he is using his model to tell us a story about country experiences. We are thus left with the impression that he may know more than he has revealed in his country studies.

The case study of Puerto Rico has relatively little descriptive analysis; it uses 1970 census data to estimate the magnitude of an income effect on fertility for households at different income levels and to provide empirical confirmation of the nonlinear relationship. The second case study shows how such a relationship can be used to interpret the Korean experience: a major redistribution of income followed by both growing income and declining fertility. This is the most thorough of the case studies, and it is perhaps no coincidence that the regression model is limited to a few pages at the end.

By far the most disturbing part of this book is the chapter entitled "Internal Policies for Income Distribution in Rural India." The data from Matar Taluka (a rural area in India's Gujarat state) are fascinating in that they resurvey an area for which data were also collected in 1930 and so permit an analysis of long-run changes in income distribution. These data are used as a point of departure for an analysis of the benefits of land reform; fertility comes up only once, in passing, when the author asserts that redistribution through land reform will have the desired income effect. The discussion of land reform which takes up fully half of this chapter is well presented and argued, but its relevance to the main theme or to anything else in this book is established tenuously at best. First, an income regression equation is estimated for Matar Taluka in which both human and nonhuman assets are included as independent variables, and it is found that variation in the nonhuman asset variables accounts for by far the largest portion of the explained variation in income. However, the author tells us that Matar Taluka is an agricultural region with large variation in the size of landholdings and that the survey year was a disastrous agricultural year for most of India;

ceteris paribus, this should lead us to expect a wider-than-usual distribution of income correlated with each farmer's harvest within the region. At the same time, we are told why the human capital variables will not predict income: because there is no variation in schooling within the region and because age pertains to the household head instead of the individual earner. Thus the empirical results for Matar Taluka may tell us what was going on during the survey year but certainly cannot justify choosing land reform over human capital development as a long-term strategy.

Repetto follows his analysis of the effects of land reform with some arguments why education cannot be effectively used to redistribute income through human capital stocks, even though that is in fact what he says has happened in Korea. His arguments often confuse inequality measures with the distribution of income. For example, since the initial distribution of schooling in Matar Taluka is very equal, the first-round impact of increasing the education level of younger cohorts will be to increase inequality of earnings, presumably *because* education improves earnings. Even ignoring potential intergenerational transfers, this is surely a most extraordinary argument against progress, and one which I doubt would impress many low-income families. In the same vein, the following paragraph contains the suggestion that since improved education would affect only new entrants to the labor force, it would be decades before their numbers could be great enough to appreciably affect the average income Then, having observed that an important value of education is that it "prepares students . . . for jobs outside the traditional sector," he says that land reform is needed because education "has more value for a future landholder than for a future agricultural laborer" (p. 149).

The irony of this analysis of the relative merits of human and nonhuman assets as sources of increased income for the poor is that although it is a small part of Repetto's book it is strategically placed at a point where policy implications are being drawn from research findings. The fact that the discussion of this point is at best superficial is especially puzzling in the light of the comparatively careful reporting of evidence in the earlier chapters. Indeed, on the whole this book contains a useful perspective on the importance of income effects on fertility and provides convincing evidence that the improvement of incomes for the poor, especially in developing countries, is necessary and even perhaps sufficient as a condition for reduced aggregate fertility.

CARMEL U. CHISWICK

University of Illinois at Chicago Circle

The Economics of the Performing Arts. By C. D. THROSBY and G. A. WITHERS. New York: St. Martin's Press, 1979. Pp. 348. $27.50.

Books on the economics of the performing arts are no longer a rarity. But this volume contains material quite different from that of its predecessors. It goes over and brings up-to-date the statistics on costs, revenues, and public support which have been analyzed before. But, in addition, it seeks to go as far as the subject allows in applying the standard techniques economists have used to analyze other industries—adapting theoretical micro models and econometric analysis to the problems of the arts.

Predictably, the success of this venture is mixed. But this is no discredit to

the authors, who are obviously masters of both the subject area and the techniques, and who display consistent good judgment on issues of public policy, which is the book's ultimate focus.

The volume deals with all of the subjects which one would expect it to consider, starting with the organizations which present performances ("the firms"), going on to consumer demand, industry structure, and then turning to public assistance to the arts.

Most subjects are treated in four different ways. Each is discussed in general terms to describe the most pressing issues to which it gives rise. Second, extensive and often illuminating descriptive statistics are provided. Third, where possible, a formal theoretical model is constructed. Finally, an econometric model is formulated and the values of its parameters are estimated.

One can quibble here and there with the more discursive portions of the book, but, by and large, they are accurate and illuminating. Of particular value is the extensive material comparing performing activity and its financing in Australia, Canada, Great Britain, New Zealand, and the United States. The value of such international comparisons has often been asserted, but in the past the materials provided for the purpose have generally been quite spotty. The major conclusions that emerge from the comparisons should, however, hardly be surprising to anyone with some acquaintance with the field. They show remarkable similarity in the economic problems faced by the arts in the different countries, in the characteristics of the audiences, in the nature of the organizations, etc. For example, it is "striking" that "proportions of the population exposed to the performing arts are substantially higher for the middle-aged, high income, high education, professional, managerial and white-collar groups" (p. 96). About the only major difference seems to be the heavy reliance on private foundations and individuals for financial support in the United States, while elsewhere such income flows almost exclusively from government. Far more remarkable are the figures on the growth of government support in the five countries studied. For the 8 years 1969 through 1976, in every one of the countries government assistance grew faster than price level, population, average hourly earnings of labor, GDP, and (with the exception of Canada) aggregate government expenditure. In Australia and the United States the degree by which government outlays on the arts outpaced these other indices was particularly remarkable. All of this is noteworthy because in the past periods of inflation have been times of financial difficulty for the arts. As a matter of fact, more recent data for the United States indicate that government outlays on the arts have continued to rise in real terms, so that while the arts have experienced financial difficulties stemming from other sources, at least so far, and with some noteworthy exceptions, real cuts in government funding have *not* been among them.

The authors' discussion of policy deals with such issues as justifications of assistance, the determination of the appropriate assistance levels and of the allocation of the funds by size of organization, art form, geographic location, etc. In discussing justifications for support the authors build their conclusions on the economists' usual grounds for government intervention: externalities, redistribution, merit goods, option demand, etc. The discussion strikes one as sensible, though no surprising conclusions emerge. They reconfirm what economists had noted before—how difficult it is to provide a very strong justification for government support of the arts in these terms except perhaps on the merit goods argument. But, then, the term "merit good" merely becomes a formal designation for the unadorned value judgment that the arts

are good for society and therefore deserve financial support. In other words, the merit good approach is not really a justification for support—it merely invents a bit of terminology to designate the desire to do so.

As already noted, the unique contribution of this book lies in its theoretical and econometric models, and these are, therefore, worth considering in a bit more detail. These models deal with a variety of subjects: consumer demand for leisure (following Buristam Linder); willingness to offer philanthropy; the theory of merit goods; propensity to provide philanthropy; and the behavior of the arts firm and its response to subsidy. The last of these is perhaps representative. Taking the nonprofit arts firms to be managed by utility maximizers constrained by a zero-loss condition, the authors proceed to formulate a utility function whose variables are attendance and quality, a production function which takes attendance to be a function of number of performances, number of distinct productions, capacity of auditorium, and quality. This is related to a cost function, whose variables are attendance and quality, and a demand function dependent on price and quality. From this, a standard partial-equilibrium diagram emerges, showing the equilibrium of the utility-maximizing firm and its relation to that of a profit maximizer. The conclusions which emerge are not particularly surprising—the utility maximizer's output (audience size) will be the larger, and its output will be increased by subsidy—even if the subsidy is a lump sum.

The econometric models also cover a wide variety of subjects: theater cost functions, consumer demand functions, voter characteristics, magnitude of government assistance, and the effects of grants on attendance. The relatively simple forms taken by the relationships used in estimation are fully explained by the complexity of the subjects and scarcity of data. More difficult to explain is the rather persistent failure of the authors to provide interesting and testable hypotheses. It is as though the statistical calculations were undertaken as ends in themselves rather than as means to get at more fundamental issues. Nevertheless, some very interesting results emerge. For example, the study of the effects of grants indicates:

> Firstly, grants to drama have been more effective at the margin in securing greater attendances than have grants to opera, ballet and music. Secondly, the marginal unit of subsidy in drama has had a relatively higher effect on attendances at low levels of grants per company, declining as the levels of subsidy increased, although this diminishing productivity at the margin is not so evident in the case of the other art forms.
>
> It must be remembered that these results obtain at the margin in a situation where the existence of grants already accounts for a major part of attendances, especially in the case of opera, ballet and music. . . . In the case of theatre, however, the wider range of levels of activity and the more profound effects of a relatively small amount of support on a small company give greater scope for grants to lead to increased attendances in the lower ranges, although at the other end of the scale (i.e., for the major subsidized companies) the marginal unit of grants is less effective in increasing attendances for reasons similar to those discussed above in relation to opera, ballet and music. [Pp. 268–69]

To summarize, this book is a very useful addition to the literature. It will prove indispensable to anyone conducting research on the subject. The au-

thors have given us up-to-date information, have contributed valuable insights, and have expanded the armory of weapons shown to be helpful. But their attempt to fit the analysis to the procrustean bed of standard economic techniques seems to have been pushed beyond the point of diminishing returns.

WILLIAM J. BAUMOL

Princeton and New York Universities

HILDA BAUMOL

Mathematica, Inc.

The Deterrent Effect of Antitrust Enforcement

Michael Kent Block and Frederick Carl Nold

Hoover Institution

Joseph Gregory Sidak

Stanford Law School

In this paper we formulate and test a model of collusive pricing in the presence of antitrust enforcement. We show that a cartel's optimal price is likely to be neither the competitive price nor the price that the cartel would set in the absence of antitrust enforcement but rather an intermediate price that depends on the levels of antitrust enforcement efforts and penalties. Our empirical results reveal that increasing antitrust enforcement in the presence of a credible threat of large damage awards has the deterrent effect of reducing markups in the bread industry.

Soon after the passage of the Sherman Act, the Supreme Court determined that horizontal minimum price fixing was so inherently injurious to consumer welfare that it should be illegal per se. Horizontal collusion has since become a major focus of federal antitrust

We wish to thank Gaetano Paderanga and Timothy Moore for valuable research assistance, and we gratefully acknowledge the helpful comments of Donald Baker, William Baxter, Gary Becker, Timothy Bresnahan, Aaron Director, George Hay, Peter Pashigian, Sam Peltzman, Mitchell Polinsky, Robert Reynolds, Kenneth Scott, George Stigler, Robert Stillman, members of the Industrial Organization Workshop at the University of Chicago, and participants in seminars at Stanford Law School, UCLA, Cornell Law School, the Rand Corporation, and the Antitrust Division, U.S. Department of Justice. This paper was prepared under grant nos. 75-NI-99-0123, 77-NI-99-0071, and 79-NI-AX-0071 from the National Institute of Law Enforcement and Criminal Justice, Law Enforcement Assistance Administration, U.S. Department of Justice. The opinions stated in this paper do not necessarily represent those of the U.S. Department of Justice.

[*Journal of Political Economy*, 1981, vol. 89, no. 3]
© 1981 by The University of Chicago 0022-3808/81/8903-0003$01.50

enforcement.[1] Through the use of criminal and civil sanctions the Department of Justice (DOJ) has sought not only to remedy specific instances of price fixing, but also to achieve general deterrence of potential price fixing.

This paper is the first systematic attempt to estimate the impact of antitrust enforcement on horizontal minimum price fixing. We develop a simple theoretical model of the collusive pricing decision and then, using data on the bread industry, assess empirically the deterrent effect of public and private antitrust enforcement on the decision to collude.

I. A Simple Model of Collusive Pricing in the Presence of Antitrust Enforcement

A. *Definitions and Assumptions*

We construct a relatively simple model to consider explicitly the effect of antitrust enforcement on the decision of firms within an industry to fix prices collusively. The model uses the following variables and definitions:

p = price of output;
$Q \equiv Q(p)$ = demand for the industry's output;
$C \equiv C(Q)$ = total cost of industry output, including a normal rate of return;
$c \equiv C(Q)/Q$ = average cost;
$mc \equiv \partial C/\partial Q$ = marginal cost;
$\lambda \equiv (p - mc)/mc$ = markup over marginal cost;
π = total profits of colluders;
γ = level of enforcement efforts directed toward detecting horizontal collusion;
$F \equiv t(p - mc)Q$ = the combined civil and criminal penalty for price fixing, where t is the anticipated damage multiple;
$d \equiv d(\lambda, \gamma)$ = the probability that a collusive pricing scheme will be detected.

The definitions of F and d are essential to our analysis and warrant further discussion.

Our specification of the penalty function, $F \equiv t(p - mc)Q$, reflects that under current statutes a price-fixing conspiracy is subject to both criminal and civil sanctions.[2] The Clayton Act's private treble dam-

[1] Posner (1970, p. 398, table 23) reports that 989 of the 1,551 Department of Justice antitrust cases between 1890 and 1969 contained charges of horizontal conspiracy.

[2] The maximum criminal sanctions for price fixing are imprisonment for 3 years, a fine of $100,000 per individual, and a fine of $1 million per corporation. We do not consider injunctive remedies in this model.

age remedy is the most formidable civil sanction and clearly relates to markups, as our specification of F reflects. If criminal sanctions are small—as they usually are—then civil sanctions provide most of the sting of antitrust enforcement, and our specification of F accurately describes the true penalty for price fixing.[3] Moreover, if, when sentencing price fixers, a district court judge considers the size of the markup to be an index of the cartel's perniciousness, then F is still reasonably descriptive on the rare occasions when criminal fines and sentences are significant relative to private treble damage awards.

In the presence of uncertain antitrust enforcement the decision to collude depends not only on the magnitude of the penalty but also on the probability of detection, which we assume increases with the markup, $[\partial d(\lambda, \gamma)]/(\partial \gamma) > 0$. The closer colluders come to the price that they would charge if no antitrust enforcement existed, the greater their chances of being discovered. Two observations support this hypothesis. First and most obvious, higher markups make customer complaints to the DOJ more likely.[4] Second, as Stigler (1968, pp. 268–70) has noted, a positive relationship probably exists between a conspiracy's detectability and its ability to prevent cheating by individual colluders. For example, a joint-sales agency would assure a set of colluders strict adherence to a monopoly price, but so visible a collusive device would be nearly assured of detection. In general, we hypothesize that the more efficiently a collusive device produces markups, the more likely it will be detected and the colluders convicted.[5] Hence, as we have assumed, the very technology of collusion makes it likely that the probability of detection increases with the markup.[6]

We also assume that the DOJ never charges noncolluding firms:

[3] A comparison of the magnitude of criminal and civil penalties appears in Sec. II. Though a substantial increase in potential criminal sanctions was legislated late in 1974 under the Antitrust Procedures and Penalties Act, there is little evidence of its effectiveness in increasing realized sanctions (see Burnham 1978).

[4] Hay and Kelley (1974) find that complaints by customers are the third most numerous method of detection among DOJ price-fixing cases.

[5] Stigler (1968) presents some indirect evidence on this point. In a sample of DOJ cases he divides the types of collusion between what he considers *efficient* and *inefficient* collusion. The average time from inception of alleged collusion to complaint is significantly shorter for the efficient forms of collusion.

[6] Our study of the bread industry revealed, as tentative empirical evidence of the relationship between the probability of detection and the markup level, that the probability of being investigated for price fixing was positively related to the markup level. We assumed that a lag exists between the price-fixing decision and its detection; then, using a one-period lagged value of our markup indicator as an explanatory variable, we estimated a logit model of the probability that DOJ would initiate an investigation for bread price fixing. The estimated intercept for our 208-observation sample was −2.49, and the estimated coefficient on the lagged-markup indicator was 5.69 with a standard error of 2.43. When the logit function is evaluated using the sample mean values, this coefficient implies that a 1-percentage-point increase in the markup increases the probability of an investigation by 5 percent. Sec. II elaborates on our specification of the markup and the nature of our sample.

$d(0, \gamma) = 0$. We assume further that DOJ does not necessarily detect a pure cartel price with certainty—that is, $d(\lambda^c, \gamma) < 1$, where λ^c is the markup that colluders would choose if no antitrust enforcement existed.

B. *Antitrust Enforcement and Optimal Collusion*

The objective of collusion is to set a price that maximizes expected cartel profits.[7] To model this price-fixing decision simply, we impose three additional assumptions. First, firms produce output under conditions of constant marginal and average cost so that $mc = c$. Second, all firms take part in the joint-pricing decision. Third, the most significant cost of any collusive device is its impact on the probability of detection. For example, the major cost of a joint-sales agency is its visibility, not its resource costs. This third assumption implies that, in the absence of antitrust enforcement, firms could duplicate perfectly the price and output levels of a monopoly.[8]

Formally, the colluders set the price p by maximizing the expected value of their objective function, $E\pi$. That is, the colluders

$$\max_p E\pi = (1 - d)\pi_1 + d\pi_2 \equiv Z, \tag{1}$$

where $\pi_1 \equiv (p - c)Q(p)$ is the profit level if the colluders avoid detection, $\pi_2 \equiv [p - c(1 + \lambda t)]Q(p)$ is the profit level if the colluders are detected, and $\lambda = (p - c)/c$ is the markup under constant costs. The necessary condition for an internal maximum is

$$Z_p = -\frac{\partial d}{\partial \lambda}\lambda t Q(p) + [Q(p) + (p - c)Q'(p)](1 - dt) = 0. \tag{2}$$

Equation (2) has a simple yet intriguing implication: If, as we have assumed, the markup significantly affects the probability of detection, then neither the competitive price ($\lambda = 0$) nor the pure cartel price markup ($\lambda = \lambda^c$) can satisfy equation (2).[9] Antitrust penalties, therefore, do not necessarily eliminate price fixing. They are, however, likely to reduce the optimal markup.[10]

[7] This assumes that the colluders are risk neutral. An analysis that allows for risk aversion is presented in Block, Nold, and Sidak (1978). The results of that analysis do not differ substantially from those presented here.

[8] As long as we restrict our attention to industries with a history of conspiracy, considering intraindustry cheating would complicate the model without significantly adding to our understanding of collusive pricing in the presence of antitrust enforcement.

[9] The pure cartel markup is the markup λ^c satisfying the condition $Q(p) + (p - c)Q'(p) = 0$.

[10] If the detection probability were not sensitive to the markup level, then the decision to price fix would be a simple either/or proposition. In this case, as long as $(1 - dt) > 0$, the colluders would set the price at the pure cartel level. Clearly no price fixing would

As long as price fixing is a favorable gamble for some markup, $(1 - dt) > 0$—and our assumptions assure it will be—then risk-neutral colluders obviously will not choose the competitive price. Also, although the "unfairness" of the price-fixing gamble at the pure cartel markup, $[1 - d(\lambda^c, \gamma)t] < 0$, is sufficient to dissuade the colluders from choosing $\lambda = \lambda^c$, it is not necessary. In general, colluders will stop marking up when both $[Q(p) + (p - c)Q'(p)]$ and $(1 - dt)$ are positive. To deter colluders from choosing the full cartel markup is rather easy; to deter all collusive pricing in a conspiracy-prone industry is nearly impossible.

C. *The Effects of Changes in Enforcement Efforts and Penalties on Collusive Markups*

To assess the effect on markups of increases in enforcement efforts, we consider the effect of changes in γ on the optimal λ, or $\partial\lambda/\partial\gamma$. Totally differentiating equation (2) with respect to λ, we obtain

$$\frac{\partial\lambda}{\partial\gamma} = \left(\frac{t}{cZ_{pp}}\right)\left\{\frac{\partial^2 d}{\partial\lambda\partial\gamma}\lambda Q + \frac{\partial d}{\partial\gamma}[Q(p) + (p - c)Q'(p)]\right\}, \qquad (3)$$

where Z_{pp} is the second-order derivative. Only in the rather perverse case where an increase in enforcement actually reduces the impact of the markup on the detection probability—that is, where $(\partial^2 d)/(\partial\lambda\partial\gamma) < 0$—does an increase in enforcement have an ambiguous effect on the markup. In most cases of practical importance we expect that DOJ efforts designed to increase the probability of detection also reduce the markup.

Increasing price-fixing penalties is often advocated as an efficient means to achieve deterrence.[11] The effect of such an increase is expressed by the comparative static derivative:

$$\frac{\partial\lambda}{\partial t} = \left(\frac{1}{cZ_{pp}}\right)\left\{\frac{\partial d}{\partial\lambda}\lambda Q + d[Q(p) + (p - c)Q'(p)]\right\}, \qquad (4)$$

which indicates that a higher penalty unambiguously decreases the optimal markup. In summary, an increase in either the probability of

occur if $(1 - dt) < 0$. Enforcement efforts would determine simply whether or not it paid for a cartel to fix prices. If it did, colluders would set markups at the full cartel level; if it did not, they would not collude. This model, in which the probability does not depend on changes in λ, would make the empirical analysis somewhat more straightforward. However, such a formulation would be less descriptive of the actual situation confronting colluders. Formally, our assumptions do not preclude a corner solution at $\lambda = \lambda^c$. If the detection probability displays a limited responsiveness to markups, then eq. (2) may not be satisfied as an equality. However, it is unlikely, given the technology of collusion and the existence of punitive sanctions for price fixing, i.e., $t > 1$, for $(1 - dt)$ to be positive at $\lambda = \lambda^c$.

[11] See Sec. III; see also U.S. Office of the President 1969, and Elzinga and Breit 1976.

detection or the penalty for price fixing can be expected in most cases of interest to reduce the collusive markup.

II. Empirical Findings

Our theoretical model suggests that increases in enforcement levels or penalties for price fixing generally reduce collusive markups. Though straightforward to derive, this implication is hardly trivial to test. Unfortunately, we do not have a set of national, or even regional, industries having identical products, costs, and demand conditions but varying levels of antitrust enforcement. We must assess the impact of antitrust enforcement in a more problematic environment.

A. Choice of Industry

To test the implications of our deterrence model, we analyzed the market for white pan bread, a homogeneous commodity not only regionally produced and consumed but also well represented among DOJ price-fixing cases. During our sample period, bread cases were the most common among DOJ's food price-fixing cases.[12] In addition, the bread industry has well-recorded annual input and output prices compiled for selected cities by the Bureau of Labor Statistics (BLS) (see U.S. Department of Labor, Bureau of Labor Statistics 1964–76, 1965–76).[13] Hence, we constructed a sample that enabled us to use both cross-sectional and temporal variations in product prices, costs, and antitrust enforcement.

B. Estimating City-specific Markups

Developing an indicator for price markups in the bread industry involved two steps. First, we examined a standard recipe for a loaf of white bread.[14] We subtracted from each bread price observation the

[12] The FTC observed that the bread-baking and distribution industry has "the essential characteristic of a conspiracy-prone industry—relatively few sellers in individual markets" (U.S. Federal Trade Commission 1967, p. 135).

[13] The BLS publications contained comparable price data for white bread for 20 major cities. For 12 of these cities the data went back as far as 1955, although for the remaining eight they extended only to 1968. Input price data, though not so complete, proved adequate to construct a sample of 228 observations: 12 major cities for 1964–76, and eight additional major cities for 1968–76. Cities included throughout the entire period (1964–76) were Baltimore, Boston, Chicago, Cleveland, Detroit, Los Angeles, New York, Philadelphia, Pittsburgh, St. Louis, San Francisco, and Washington, D.C. For 1966–76, we were able to add Atlanta, Cincinnati, Dallas, Houston, Kansas City, Minneapolis, San Diego, and Seattle to the sample.

[14] The cost of ingredients (IC) for a 1-pound loaf of white bread is $IC = .6350P_F + .0571P_S + .0026P_O + .035P_M$, where P_F is price/pound of flour, P_S is price/pound of

TABLE 1

EFFECT OF NONINGREDIENT INPUT COSTS ON ADJUSTED BREAD PRICE (PADJ)

Independent Variables	
PELEC	.256*
	(4.62)†
PGAS	2.49
	(3.15)
LABOR	1.07
	(7.14)
Constant	7.05
Observations (N)	228
R^2	.49
F-statistic (3,224)	72.93

* The estimated coefficient
† The value of the estimated coefficient divided by its estimated standard error

component attributable to the cost of the ingredients, IC, calculated from the recipe and from BLS ingredient prices. Formally, we defined the recipe-adjusted bread price, PADJ, as:

$$\mathrm{PADJ}_{it} \equiv p_{it} - \mathrm{IC}_{it}, \tag{5}$$

where i is the city and t the time index.[15] Differences in profits and other noningredient input costs cause PADJ to vary across cities and over time.

As a second step in developing a measure of markups, we estimated the amount of variation in PADJ attributable to variations in energy and labor costs. Table 1 contains the results of this analysis; PELEC and PGAS are measures of electricity and natural gas prices, and LABOR is a measure of wage rates for truck drivers. We assume this wage to be a proxy for general labor costs.[16] The results in table 1 suggest that variations in energy and labor input costs account for a reasonable amount of the variation in adjusted bread prices.[17]

sugar, P_O the price/pound of cooking oil, and P_M is the price/pound of dry milk. All prices are retail, except dry milk, which is wholesale. Cooking oil is used as a proxy for shortening (U.S. Executive Office of the President 1977, p. 7).

[15] The sources of price data were U.S. Department of Labor, Bureau of Labor Statistics (1965–76; 1969).

[16] Specifications using food and kindred workers' wages as well as alternative functional forms were also estimated. These different specifications yielded similar results (see Block et al. 1978).

[17] Of course, measures of goodness of fit for this regression substantially understate the importance of variations in input prices since the dependent variable, PADJ, is already net of the recipe or ingredient. Our entire procedure, taking account of both IC and the estimated cost function, accounts for approximately 82 percent of the variation in the price of bread (p). In general, prices for all significant inputs are included in cost functions. Our asymmetric treatment of input prices amounts to using

To construct a measure of the markup on bread, we then used the recipe costs and the results of this regression to estimate M, the markup indicator:

$$M \equiv [p - (\mathrm{NIC} + \mathrm{IC})]/(\mathrm{NIC} + \mathrm{IC}), \tag{6}$$

where all subscripts are suppressed and NIC refers to the fitted values generated by the equation in table 1.[18] Equation (6) estimates the markup by first subtracting from the market price the sum of the known ingredient costs (IC) and an estimate of other noningredient costs (NIC), and then dividing this difference by estimated unit costs. The higher the residual as a proportion of unit costs, the higher the markup.

C. *Estimating the Effect of DOJ Antitrust Enforcement on Collusive Markups*

Antitrust enforcement and penalties are only two of the many factors that actually determine markup levels.[19] Certainly, all factors that influence the price elasticity of demand also influence the markup level. Moreover, although we have assumed antitrust penalties to be the only cost of collusion, the resource cost of managing and policing a cartel also influences the optimal markup in the presence of antitrust enforcement. Assessing the effect of antitrust enforcement, therefore, requires a workable method of controlling for such outside influences. Influences generated by aspects of market structure that do not change rapidly—or at least that change significantly less rapidly than the antitrust variables—can be controlled by considering not the markup level but *changes* in the markup level.[20] We therefore used first differences in the markup level, or

$$\Delta M_{it} = M_{it} - M_{it-1}, \tag{7}$$

in testing the effectiveness of antitrust enforcement.[21]

---

a priori information—the recipe for bread—to supplant estimation of some coefficients. It may be argued that this procedure increases the efficiency of the estimation, but it precludes substitution among raw materials and other inputs.

[18] Because a simple regression generates NIC, approximately half the observations for NIC exceed PADJ and thus generate a negative M. For our purposes, these negative values pose no problems.

[19] We structured our test of antitrust enforcement around markups rather than the traditional—but indirect—test of overall profitability or rate of return. One reason for using the markup was the very directness of the test. Also, the availability of regional and city data enabled us to conduct more powerful tests of the effectiveness of antitrust enforcement than could have been performed with available data on rates of return.

[20] Controlling for these factors by actually assembling the relevant information on market structure for this industry at the city or SMSA level was not feasible.

[21] This procedure also facilitated our use below of a straightforward measure of regional antitrust enforcement.

1. Measuring Detection Probabilities, DOJ Enforcement Efforts, and Remedial Effects

We assume in our formal model of collusion that for any markup level the probability of detection is a function of DOJ enforcement efforts. Unfortunately, this straightforward theoretical proposition does not suggest a unique empirical counterpart for measuring either the probability of detection or DOJ's enforcement efforts. This formulation does suggest, however, that for any markup level the probability of detection is related to the capacity of the Antitrust Division to initiate cases. To the extent that the Division's litigation capacity relates to its expenditure level, the Division's annual budget should provide an indirect measure of this enforcement capability.[22]

A more direct measure of enforcement is simply DOJ price-fixing prosecutions. A price-fixing prosecution is rare enough for colluding firms in the affected industry to take special notice.[23] We have assumed, therefore, that each price-fixing case that the DOJ brings against a bread producer increases, for any markup level, the probability of prosecution perceived by other bread producers operating in the same DOJ region.[24]

We had sufficient price data to estimate changes in markups for 1965–76. We constructed a regional antitrust enforcement variable, DOJREG, for this period by setting the variable equal to one for each city within a region where the Antitrust Division filed an action that year—except for the city incurring the action—and by setting the variable equal to zero otherwise.[25] In other words, DOJREG is a shift

[22] A positive and significant relationship exists between the Antitrust Division's annual budget and the number of price-fixing cases brought by the Division. Using data on price-fixing cases during 1964–76 supplied by the Antitrust Division's Economic Policy Office, we estimated the relationship:

$$\text{DOJPF} = -11.39 + \underset{(2.42)}{.003}\text{BUDGET},$$

between the number of price-fixing cases brought annually by the Division (DOJPF) and its budget (BUDGET) measured in thousands of 1967 dollars (2.42 is the t-statistic).

[23] We can provide a formal rationale for this observation by assuming that colluders use Bayesian methods to estimate the probability that they will be apprehended in a particular period. In this formulation, whenever colluders are apprehended, colluders' estimate of the probability of apprehension increases, and that increase is dramatic if their a priori distribution is diffuse and has a small mean. Also, after an initial impact, the effect of the case on the apprehension probability estimated by the colluders deteriorates if no new cases are brought.

[24] The Antitrust Division operated seven regional offices during the years of our sample: New York, Philadelphia, Cleveland, Chicago, San Francisco, Los Angeles, and Atlanta. The Antitrust Division established a Dallas office in 1976 and realigned the areas covered by the regional offices. This study used the earlier seven regions for 1976 rather than realigning all regions for that single year. For coverage of regions, see U.S. Department of Justice, Antitrust Division (1978).

[25] This variable was constructed using data from the Commerce Clearing House (1955–75; 1966–76). A summary of those data appears in Block et al. (1978). Filing dates and other summary information were checked against a special listing of bread

variable designed to capture the changes in the perceived probabilities of cartel failure. Its form presumes that the colluders' estimate of the probability that DOJ will initiate a case for collusion increases after a DOJ price-fixing action within the same region.[26]

The variable DOJREM measures the effect of the antitrust action in the city where the Antitrust Division actually prosecuted a bread producer. In constructing DOJREM, we assumed that the impact of the antitrust action on the firms specifically prosecuted would differ significantly from the impact on firms in other cities in the same region. Basically we assumed that, for strategic reasons, the timing of the reduction in markups by prosecuted firms would not coincide with the reduction made by firms in nonaffected cities. To capture this effect, we set DOJREM equal to one in a city 1 year after the Antitrust Division had filed an action there.

2. Estimated Deterrent Effects

In table 2 we present our estimates of the effect of DOJ enforcement variables on markups in the bread industry. Again, the dependent variable, ΔM, is the annual change in the markup on white bread for the cities in our sample. The results in table 2 strongly suggest a deterrent effect of DOJ enforcement efforts.[27] First, the coefficient on

price-fixing cases prepared for us by the Economic Policy Office, Antitrust Division, U.S. Department of Justice, for the years 1963–76.

[26] An alternative hypothesis would be that, since DOJ has limited resources, a prosecution in a region signals that an additional prosecution in the same region is unlikely. Hence a firm actually would reduce its estimated probability of detection after a DOJ case in its own region. This hypothesis is equivalent to predicting that drivers on a highway accelerate when they see a car already being ticketed.

[27] Our formulation, which posits that increases in the markup increase the detection probability, generates the structural equations: $\Delta d = \xi + \beta\Delta\lambda + \tau\Delta\text{BUDGET} + \eta\text{DOJREG}$ and $\Delta\lambda = \psi + \alpha\Delta d + \theta\text{DOJREM}$. Therefore, the reduced forms are

$$\Delta d = \omega + \frac{\tau}{1 - \alpha\beta}\Delta\text{BUDGET} + \frac{\eta}{1 - \alpha\beta}\text{DOJREG} + \frac{\beta\theta}{1 - \alpha\beta}\text{DOJREM}$$

and

$$\Delta\lambda = \rho + \frac{\alpha\tau}{1 - \alpha\beta}\Delta\text{BUDGET} + \frac{\alpha\eta}{1 - \alpha\beta}\text{DOJREG} + \frac{\theta}{1 - \alpha\beta}\text{DOJREM}.$$

Our estimated reduced-form coefficients in table 2 are of the form $(\alpha\tau)/(1 - \alpha\beta)$, $(\alpha\eta)/(1 - \alpha\beta)$, and $(\theta)/(1 - \alpha\beta)$. We are mainly interested in deducing the sign of α given the sign of the reduced-form coefficients. The usual approach would be to estimate the other reduced-form equation in the system which takes Δd to be a function of the exogenous variable. Unfortunately this approach cannot be applied directly to this problem. We do not have a proxy for Δd or measures of the detection probability at a regional level. Even at a national level measuring d is problematic. However, given our formulation of the determinants of detection, if we assume that DOJ expenditures have some efficacy—i.e., $\tau > 0$—then the condition necessary to deduce the nonpositivity of α from the reduced-form coefficients is that $(1 - \alpha\beta) > 0$. This condition is required for Walrasian stability in the collusion market and in conjunction with our reduced-form

TABLE 2

ESTIMATED EFFECTS OF CHANGES IN DOJ ENFORCEMENT ON CHANGES IN MARKUPS IN THE BREAD INDUSTRY, 1965–76

Independent Variables				
ΔBUDGET	−.015*	−.015	−.024	−.020
	(−2.74)†	(−2.68)	(−4.06)	(−3.65)
DOJREG	−.025	−.026	−.025	−.027
	(−2.05)	(−2.21)	(−2.09)	(−2.26)
DOJREM		−.046	−.046	−.044
		(−2.32)	(−2.41)	(−2.32)
ΔFOODM			+.058	...
			(2.33)	
ΔGENM				−.010
				(−1.60)
Constant	.011	.013	.014	.017
R^2	.055	.082	.113	.101
F-statistic	5.93 (2,205)	6.04 (3,204)	6.47 (4,203)	5.68 (4,203)

NOTE.—Each regression is based on 208 observations.
* This coefficient is estimated per million dollars
† The value of the estimated coefficient divided by its estimated standard error

our general measure of changes in enforcement capacity—the change in the real value of the Antitrust Division's budget, ΔBUDGET—is negative and significant. In other words, an increase in the enforcement capacity of the Antitrust Division appears to reduce markups on white bread. Second, the coefficient on our direct measure of DOJ's enforcement activity (DOJREG) is negative and significant, suggesting that a price-fixing case against bakers in one city induces bakers in neighboring cities to reduce markups. This result comports both with our formal theoretical results and with the related conjecture in the Stigler report that "every victory" in seeking out price fixing "weakens the efficiency of undetected collusion."[28] Finally, the coefficient on DOJREM, the variable measuring the remedial effect of a price-fixing case, is negative and significant. Once discovered and prosecuted, colluders apparently "remedy" their price fixing by reducing their markups in the following year.[29]

results implies that there is a deterrent effect, i.e., $\alpha < 0$. In addition to this theoretical restriction on the sign of $(1 - \alpha\beta)$, we noted previously that there is a positive relationship between the total number of price-fixing cases brought by the Antitrust Division and the Division's budget level. If we can assume that other collusion-prone industries made decisions about collusion in the same way as bread producers, then this aggregate regression gives information about the second set of reduced-form coefficients. Again, if $\tau > 0$, the sign of the reduced-form coefficient on ΔBUDGET indicates that $(1 - \alpha\beta)$ is positive.

[28] See U.S. Office of the President, President Nixon's Task Force on Productivity and Competition 1969.

[29] It is only necessary to assume that DOJ capacity is productive ($\tau > 0$) to infer from our empirical results that $\eta > 0$, or that filing of a DOJ case in the region increases d.

The ΔFOODM and ΔGENM variables in table 2 control for general year-to-year variations in manufacturing markups that might not be adequately controlled by the first-difference procedure. The variable ΔFOODM is an annual series of the first differences in markups of food-and-kindred-product manufacturers; ΔGENM is an analogous series for all manufacturing firms.[30] The coefficient on ΔFOODM is of the expected sign, suggesting that, holding enforcement constant, markups in the bread industry move in the same direction as markups in other food-related industries. Although using the general manufacturing markup (ΔGENM) in the equation does not alter the estimates of the coefficients on the enforcement variables, the sign of the coefficient on ΔGENM is curious.

D. *Class Actions and the Effect of Antitrust Enforcement on Markups*

Historically, trial judges have punished price fixers leniently.[31] The cases collected for this study were no exception. All but two of the 17 bread price-fixing cases between 1957 and 1976 involved *nolo contendere* pleas, and in only one case did a defendant serve an actual prison sentence; moreover, total fines as a percentage of the pretax profits of the colluding firms averaged only 7 percent.[32] Neither imprisonment nor monetary penalties posed a credible threat to colluding firms. We hypothesized, therefore, that the deterrent effect of DOJ's enforcement efforts came not from the threat of publicly imposed fines or imprisonment, but from the increased likelihood of an award of private treble damages to bread consumers or distributors.[33]

[30] Both series were computed from data reported by the Federal Trade Commission in *Quarterly Financial Report for Manufacturing, Mining, and Trade Corporations*, in U.S. Council of Economic Advisers, *Economic Report of the President* (annual). Of course, both markup measures are influenced by antitrust enforcement efforts to the extent that collusion is important in these manufacturing industries.

[31] E.g., during 1966–76 over 85 percent of all price fixers who were convicted or who pleaded *nolo contendere* did not serve prison terms (see Block et al. 1978).

[32] For the 16 criminal cases involving either a *nolo* plea or a conviction the actual criminal fines for price fixing were:

	Minimum	Maximum	Average
Individuals ($)	1,031	9,966	4,025
Firms ($)	2,930	50,638	20,690
Fines, as % of defendants' annual sales	.09	1.0	.31

All fines are in 1976 dollars. Data on the ratio of fines to sales were derived from court documents relating to 13 cases (76 firms) during 1957–75. The overall ratio of net profits to sales was obtained from a study by the U.S. Executive Office of the President, Council on Wage and Price Stability (1977).

[33] Civil actions alleging horizontal price fixing are possible but, certainly in this industry, uncommon without a preceding criminal case.

Since bread price fixing generally causes a small injury to many individual distributors and consumers, private damage recovery is usually feasible only through a class-action suit. Class actions enable plaintiffs who are numerous, and whose independent damage claims are too insignificant to justify litigation, to maintain a single action for their aggregate damages.[34] District court documents revealed that settlements in class actions for price fixing in the bread industry were almost 10 times greater than government-imposed fines.[35]

To test the hypothesis that private class actions actually provided the effective penalty in price-fixing cases, we partitioned our sample into the periods before and after class actions became a credible threat in the bread industry. Since only one class action in our sample did not follow a DOJ case, we assumed in partitioning the sample that

[34] These actions, brought under Rule 23 of the Federal Rules of Civil Procedure, have become far more frequent since the Supreme Court amended that rule in 1966. The amended Rule 23 contributed to this greater frequency by changing the procedure for becoming a member of the class. Originally, Rule 23 required persons to "opt into" the class before they could benefit from the adjudication of the class's cause of action; the amended rule instead presumed persons to be class members unless they affirmatively "opted out" of the class. Although surprisingly few historical data exist on class actions, the data that do exist—particularly the information on docket entries for the Southern District of New York collected by the American College of Trial Lawyers' Special Committee on Rule 23—suggest that the 1966 amendments to Rule 23 made the class action a much more attractive legal device (American College of Trial Lawyers 1972). A recent alternative to the antitrust consumer class action is the *parens patriae* device by which a state attorney general may sue on behalf of the consumers in his state. For the purposes of deterrence, *parens patriae* actions are virtually identical to class actions (see Block et al. 1978).

[35] A search of Commerce Clearing House (1955–75; 1966–76), McLaughlin (1976), Newberg (1977), and the LEXIS computer file (1978) revealed seven major class actions for price fixing in the bread industry since 1966. According to district court documents, class-action settlements or damages during 1971–76 obtained for price fixing in the bread industry were.

	Minimum	Maximum	Average
Cases ($)	1,197,810	6,100,000	1,998,646
Firms ($)	39,562	1,220,000	293,919
Damages as % of defendants' annual sales	.41	19.68	2.87

All fines are in 1976 dollars. These figures are based on data for five of the seven recorded cases. One of the seven original class actions had no settlement or award since the district court refused to certify the class. No data at all were available on another case. All five cases were settled rather than litigated to judgment. Details of these class actions appear in Block et al. (1978, appendix table VI). Whereas average damages as a percentage of defendants' sales were only 10 times average fines as a percentage of defendants' sales, average damages per firm were almost 15 times average fines per firm. This occurred because the average sales per defendant differed significantly between the two samples: $5,795,271 for the fine calculations and $10,254,205 in the class-action calculations.

TABLE 3

ESTIMATED EFFECT OF DOJ ENFORCEMENT AND CLASS ACTIONS ON MARKUPS IN THE BREAD INDUSTRY, 1965–76

Independent Variables		
ΔBUDGET11	−.018*	...
	(−.622)†	
ΔBUDGET12	−.014	...
	(−2.68)	
ΔBUDGET21	...	−.002
		(−.158)
ΔBUDGET22	..	−.019
		(−3.53)
DOJREG11	+.004	...
	(.160)	
DOJREG12	−.037	...
	(−2.70)	
DOJREG21	...	−.019
		(−1.03)
DOJREG22	. .	−.029
		(−1.87)
DOJREM	−.046	−.042
	(−2.43)	(−2.19)
Constant	.013	013
Observations (N)	208	208
R^2	.0926	.1024
F-statistic	4.13 (5,202)	4.60 (5,202)

* This coefficient is estimated per million dollars

† The value of the estimated coefficient divided by its estimated standard error

class actions affected primarily the penalty cost of detection, not the probability of detection.[36] We considered as partition dates the year that a district court first certified a class in a bread case (1970) and the year that the Administrative Office of the Courts began reporting class-action activity (1972). Before 1970 class actions probably were not a credible threat in the bread industry; by 1972 no doubt remained that class actions had become prevalent and important. To accomplish the actual partitioning of the sample, we split the enforcement variables ΔBUDGET and DOJREG around 1970 and 1972. In table 3, ΔBUDGET11 and DOJREG11 are ΔBUDGET and DOJREG for 1965–69, while ΔBUDGET12 and DOJREG12 are the variables for 1970–76. The comparable variables partitioned around

[36] To determine when, after the 1966 amendments to the class-action provisions in the Federal Rules of Civil Procedure, the class action became a credible threat in the bread industry, we searched for both recorded and unrecorded class actions involving horizontal price fixing of bread products. None of the seven class actions we found was filed before 1968, actually certified before 1970, or settled before 1971.

1972 instead of 1970 are ΔBUDGET21, ΔBUDGET22, DOJREG21, DOJREG22.[37]

The estimates in table 3 are consistent with our hypothesis that class actions represent the effective penalty in price-fixing cases. This result does not seem to depend upon whether we partition the enforcement variables around the date of the first class certification or around the date of the first reporting of class-action activity by the Administrative Office of the Courts. For either partition, only in the latter period, when class actions represented a credible threat, did a significant deterrent effect result from either an increase in the Antitrust Division's resources or from the actual prosection of a horizontal price-fixing conspiracy.[38]

The pattern in the bread industry is clear. A successful federal prosecution signals to consumers that a treble damage suit has become feasible. Private plaintiffs subsequently provide the effective penalty in the form of class actions for treble damages.[39] For price fixing in this conspiracy-prone industry—and, we suspect, for price fixing in general—deterrence has been a product of both public and private enforcement efforts.[40]

[37] Formally, the definitions of the variables are: ΔBUDGET21 = ΔBUDGET, 1966–71, zero otherwise; ΔBUDGET22 = ΔBUDGET, 1972–76, zero otherwise; DOJREG21 = DOJREG, 1966–71, zero otherwise; DOJREG22 = DOJREG, 1972–76, zero otherwise.

[38] While the estimates in table 3 do not include the markup controls, ΔFOODM and ΔGENM, regressions with these controls yielded similar results. Several factors possibly confound this analysis of the deterrent effect of class actions. First, the average criminal fine imposed by district courts might have increased over the period, or might simply have been larger when class actions became a relevant concern. To check for this possibly confounding influence, we estimated time trends for several measures of criminal fines. Overall, the evidence suggests that an increase in criminal penalties did not confound our results on the deterrent effect of class-action suits. Second, the early 1970s included a period of price controls, and DOJREG12 and/or ΔBUDGET12 possibly proxied for the effect of these controls. We tested whether the enforcement variables performed this role by including directly in the markup regression a dummy variable for price controls. The results of this procedure suggest that, although price controls had a significant depressing effect on markups, the enforcement variables were not merely a proxy for price controls. Finally, we controlled for pure time trends associated with changes in demand or market structure within each city. This was accomplished by regressing the dependent variable (ΔM) against city dummies along with the deterrent variable. The effect of the enforcement variables remained significant and unchanged and nearly all of the city dummies were statistically insignificant (see Block et al. 1978).

[39] A *nolo* plea apparently is a sufficient signal. Although guilty verdicts are prima facie evidence in a treble damage suit and should induce private enforcement, they are rare in price-fixing cases. In our sample of 17 price-fixing cases, one case ended in an acquittal and another in a conviction; the other 15 involved *nolo* pleas. Yet even a government case that ends in a *nolo* plea signals to potential private plaintiffs that they probably would prevail in a damage suit against the alleged price-fixing conspiracy. In fact, all the reported class actions in our sample that followed a government action and received class certification eventually obtained a settlement award

[40] The amendment of Rule 23 in 1966, of course, was the event which facilitated class actions, thereby increasing the power of the federal government to deter price fixing.

III. Conclusion

In this paper we formulated and tested a simple model of collusive pricing in the presence of antitrust enforcement. We showed that if a cartel's probability of detection increases with its markup, then the cartel's optimal price is neither the competitive price nor, in most cases, the price that a cartel would charge in the absence of antitrust enforcement, but rather an intermediate price that depends on the levels of antitrust enforcement efforts and penalties.

Our empirical results revealed that increasing DOJ's enforcement capacity or filing a DOJ price-fixing complaint had the deterrent effect of reducing markups in the bread industry. We noted that government-imposed price-fixing penalties were trivial and found support for the proposition that the effective deterrent to price fixing was the credible threat of large damage awards to private class actions that followed DOJ's case against the same conspiracy. Consequently, only after class actions became a credible private remedy did the Antitrust Division's enforcement capacity or its filing of a bread price-fixing case deter collusion in the conspiracy-prone bread industry.

References

American College of Trial Lawyers. *Report and Recommendations of the Special Committee on Rule 23 of the Federal Rules of Civil Procedure*. Los Angeles: American Coll. Trial Lawyers, 1972.

Block, Michael K.; Nold, Frederick C.; and Sidak, J. Gregory. "The Deterrent Effect of Antitrust Enforcement: A Theoretical and Empirical Analysis." Technical Report no. ISDDE-1-78, Center for Econometric Studies of the Justice System, Hoover Institution, Stanford Univ., December 1978.

Burnham, David. "Tougher Penalties in Antitrust Cases." *New York Times* (March 16, 1978).

Commerce Clearing House. *Trade Regulation Reporter*. New York: Commerce Clearing House, various issues, 1955–75.

———. *Trade Cases*. New York: Commerce Clearing House, various issues, 1966–76.

Elzinga, Kenneth G., and Breit, William. *The Antitrust Penalties: A Study in Law and Economics*. New Haven, Conn.: Yale Univ. Press, 1976.

Hay, George A., and Kelley, Daniel. "An Empirical Survey of Price Fixing Conspiracies." *J. Law and Econ.* 17 (April 1974): 13–38.

McLaughlin, Joseph T., ed. *Federal Class Action Digest 1976*. New York: Practising Law Inst., 1976.

Landmark Supreme Court decisions regarding class-action suits such as Eisen v. Carlisle and Jacquelin (417 U.S. 156 [1974]) and legislation such as the Hart-Scott-Rodino Antitrust Improvement Acts are likely to have similarly far-reaching effects. Some very preliminary attempts to assess the impact of these events on collusion appear in Block et al. (1978).

Newberg, Herbert B. *Newberg on Class Actions: A Manual for Group Litigation at the Federal and State Levels.* New York: McGraw-Hill, 1977.

Posner, Richard A. "A Statistical Study of Antitrust Enforcement." *J. Law and Econ.* 13 (October 1970): 365–419.

Stigler, George J. *The Organization of Industry.* Homewood, Ill.: Irwin, 1968.

U.S. Council of Economic Advisers. *Economic Report of the President.* Washington: Government Printing Office, annual.

U.S. Department of Justice, Antitrust Division. "Memorandum to All Personnel, Re: Organization and Operation of the Antitrust Division." Directive no. 10-73, September 30, 1973.

U.S. Department of Labor, Bureau of Labor Statistics. *Wholesale Prices and Price Indexes.* Washington: Bur. Labor Statis., various issues, 1964–76.

———. *Estimated Retail Food Prices by City* Washington: Bur. Labor Statis., various issues, 1965–76.

———. *Retail Prices of Food, 1964–68* Bulletin no. 1632. Washington: Bur. Labor Statis., 1969.

U.S. Executive Office of the President, Council on Wage and Price Stability. *A Study of Bread Prices.* Washington: Executive Office of the President, April 1977.

U.S. Federal Trade Commission. *Economic Report on the Baking Industry.* Washington: Federal Trade Commission, 1967.

U.S. Office of the President, President Nixon's Task Force on Productivity and Competition. *Recommended Changes in Antitrust Policies.* Washington: Government Printing Office, 1969.

The Market Evaluation of Human Capital: The Case of Indentured Servitude

David W. Galenson

University of Chicago and California Institute of Technology

This paper examines the market for human capital created by the institution of indentured servitude in colonial America. The indenture system allowed English emigrants to obtain passage to the colonies by selling claims on their future labor. With the size of the debt approximately equal for all emigrants, the length of the term for which a servant was bound is predicted to have varied inversely with expected productivity in the colonies. Analysis of two collections of contracts made in the seventeenth and eighteenth centuries supports the prediction. Age, skill, and literacy were negatively related to length of indenture. Women received shorter terms than men at young ages, while servants bound for the West Indies and those bound in periods of high colonial demand for labor also received reductions.

I. Introduction

During the colonial period of American history, two institutions existed which provided for the explicit valuation of stocks of human capital in the market. One of these was slavery, under which blacks and their progeny were held in service for life. The other was indentured servitude, under which whites were bound to service for limited periods of time.

I am grateful to Stanley Engerman for discussions of many of the issues treated in this paper and comments on an earlier draft. I would also like to thank Andrew Abel, Gary Becker, Lance Davis, Robert Fogel, Richard Freeman, Russell Menard, Frederic Mishkin, Melvin Reder, Sherwin Rosen, T. W. Schultz, George Stigler, and participants in seminars at Columbia University, the University of Chicago, the University of Illinois, the Newberry Library, the University of Southampton, and the 1980 Cliometrics Conference for their suggestions and comments.

Journal of Political Economy, 1981, vol. 89, no. 3]
© 1981 by The University of Chicago. 0022-3808/81/8903-0007$01.50

As a result of both its direct importance for the history of the nineteenth century and its indirect importance for the history of the twentieth, slavery has long been one of the central concerns of American social and economic historians. However, in the labor markets of many regions, for substantial periods during the first half of the colonial era, indentured servitude was of greater quantitative importance than slavery.[1] In the eighteenth century the quantitative significance of indentured servitude declined in most of these areas, as slaves were substituted for servants in the sugar fields of the West Indies, in the tobacco fields of the Chesapeake colonies, and in the rice fields of South Carolina.[2] Yet indentured servitude nonetheless continued to perform an important role in the colonial labor market by providing skilled craftsmen and managers to the large plantations of the West Indies and the southern mainland colonies.[3]

The essential difference between servitude and slavery was that it was the labor of the servant, rather than the person, which was bought and sold. Because less attention has been devoted to the study of indentured servitude, we are less familiar with the history of the institution. An economic investigation of its functioning, based on the analysis of quantitative data generated by the operation of the market for servants, can therefore serve a dual purpose: It can provide evidence of how capital values were established when claims on the labor of humans for long periods were traded in the market, while at the same time increasing our understanding of the economic basis of white servitude in colonial America.[4]

II. The Indenture Bargain

Indentured servitude was a credit system under which human labor was leased. It functioned through two markets linked by a recruiting agent. In England, in the first market, a prospective servant signed a

[1] An example is Maryland, where a study of probate inventories has indicated that as late as 1674–79 the ratio of servants to slaves held in estates was 3.88 (Menard 1977*b*, p. 360). It should be pointed out that in some areas neither type of bound labor was quantitatively important. Slaves never accounted for more than 3 percent of New England's total population in the colonial period, and the share of servants was probably of a similar magnitude (Greene 1942, chap. 3; Abbot Emerson Smith 1947, pp. 28–29). A generalization that held for a number of major colonial regions is that the share of bound workers in a region's labor force tended to rise as one traveled south from New England, with progressively increasing shares in the middle colonies, the Chesapeake, South Carolina, and the West Indies. For evidence on population composition, see Greene and Harrington (1932), Sutherland (1936), and Wells (1975).

[2] For accounts of this process in the West Indies, see Dunn (1973); for the Chesapeake, see Menard (1977*b*); for South Carolina, see Wood (1975).

[3] On this change, see Gray (1958, 1:350) and Pares (1960, p 19); for additional evidence see Galenson (1979*a*, chap. 10).

[4] The present paper extends the analysis and empirical findings of Galenson (1977*b*).

contract, or "indenture," with a merchant, promising to serve the latter or his assignees in a particular colony for a given period under stated conditions. The servant was then transported to the specified colonial destination, where the merchant or his representative sold his contract to a colonial planter or farmer in the second market. In return for the commitment of his labor, the servant received passage to the designated colony, maintenance during the term of the contract, and certain freedom dues at its conclusion. Once signed, the indenture was negotiable property, and at any time before its conclusion the servant could be sold to a new master for the balance of his term. When the contract expired, the servant became free. The conditions of servitude were regulated by colonial statutes as well as by agreements written into the contracts. The terms of the contract were binding upon both master and servant.

The indenture system normally operated within a context of competitive markets both in England and in the colonies. Servants were one important available backhaul cargo for English ships engaged in the trade for colonial sugar, tobacco, rice, and other agricultural staples.[5] Contemporary fare quotations indicate that the charges for passage from England to America were uniform at a given time for all individuals and did not vary by specific colonial destination.[6]

The planter's demand for indentured servants was based on his calculation of the discounted value of their net future earnings, after deducting the expected costs of the servant to him. The present value of the servant to the planter therefore depended upon the expected value of the servant's output in each year of the contract; the expected cost of maintenance, supervision, and training for the servant during each year of the term; the discount rate; and the value of the freedom dues to be paid to the servant. These, or analogous variables, are the same considerations which enter into the derivation of a free worker's net age-wealth profile (Becker 1975, p. 223). Yet one critical difference is that, whereas for free workers evidence on flows is used to calculate the values of capital stocks, in the case of servants these calculations were performed by planters, who based their demand for stocks of bound labor for fixed terms on their calculations involving

[5] See Bruce 1907, 1.622; Smith 1947, p. 39; and Middleton 1953, pp. 145–56.

[6] While £6 was the fare cited in the early colonial period, after the middle of the seventeenth century £5 was the fare normally quoted for passage to all colonies. For references to quotations of passage charges from England to a number of colonies, see, e.g., John Smith (1624, pt. 1, p. 162), Purchas (1625, pt. 4, p. 1791); Bullock (1649, p. 47); Taunton (January 4, 1670); Wilson (1682, p. 19); Littleton (1689, p. 17); Jeaffreson (1878, pt. 2, p. 102); Kingsbury (1906, pt. 1, pp. 277–78); Abbot Emerson Smith (1947, p. 35); and Alexander (1972, p. 45). A qualification to the statement in the text is possible seasonal variation in the cost of delivering servants to some colonies, discussed below.

the relevant flows. Therefore, like the slave market, the market for indentured labor produced capital values, and the flows underlying these must be inferred.

The institutional arrangements which provided for the ownership of human capital for discrete periods produced potential differences between the patterns of human-capital values under indentured servitude and those implicit in the operation of free labor markets. An advantage of the indenture system for the planter, relative to the hiring of free workers, was the greater control it gave him over the servant's time and effort once the bargain was made. Because the servant would have been compensated for his loss of freedom, in the absence of uncertainty, the price of indentured human capital would have been higher than that of free workers.[7] However, in practice, uncertainty makes the direction of the inequality between these prices unpredictable without additional information, for it cannot be determined a priori whether the insurance value of the contracts was normally greater to the planter or the servant.[8]

All servants who migrated to America incurred debts of similar value. As noted above, passage charges were uniform for all servants, and maintenance costs and freedom dues varied little across individuals. As a result, every servant contract was a promise to repay approximately the same sum of money. Therefore, the present discounted value of every servant's contract should have been approximately the same at the time of binding. Since the productivity of servants varied, the conditions of their indentures had to vary. The higher ther servant's marginal-value product above his maintenance, the faster he could repay the loan made to him, and the shorter the term of the contract. The length of indenture across servants should, therefore, have been negatively correlated with individual productivity or, equivalently, with the market valuation of the current flow of income generated by the individual's stock of human capital.

[7] This assumes a solution to the problem of monitoring the servant's work in which the productivity gains from the master's control over the servant's time and effort were not offset by shirking.

[8] In a world in which futures contracts for free wage labor could be made with certain fulfillment, the present value of an indenture for a given number of years and of a series of contracts for hires for the same years would differ only by a premium which would reimburse the servant for the loss of freedom resulting from his residence in the master's household and due to other legal provisions governing servitude. However, in the absence of these guaranteed futures contracts, in some cases the master might also have been willing to pay more for an indenture because of the assurance it gave him of labor supply in peak seasons or future years; similarly, workers might sometimes have been willing to accept lower implicit wages in return for the guarantee of employment the indenture represented. The relative cost of this insurance to master and servant could vary, and as a result the relation between the implicit wage paid to servants and the hire rate for free workers is indeterminate.

Strictly, if costs had been precisely the same for all servants, all should have been bound on conditions that would yield the same expected price on arrival in the colonies. This, of course, does not imply that no variation should have occurred in the realized auction prices of servants in the colonies, for random disturbances between the time of binding and auction—such as illnesses of servants during the voyage or changes in the price of colonial outputs—could have produced differences between actual and expected prices. What the analysis does imply, however, is that if the market for contracts of servitude was efficient, variation in servants' auction prices should have been uncorrelated with all the characteristics of the servants which were known at the time of making the indenture bargain.[9]

Very little evidence of the prices of newly arrived servants at colonial auctions has survived: From the entire colonial period the only known records of auctions of English servants come from the accounts of two ships, the *Tristam and Jeane,* which sold 68 servants in Virginia in late 1636 or 1637, and the *Abraham*, which sold 56 servants in Barbados in January 1637 (London, Public Record Office, 1636, 1637). The listings of the amounts paid for the servants show that the median and modal prices of both men and women were identical—500 pounds of tobacco—in both auctions. Thus, although as will be seen both sex and colonial destination had a significant effect on the length of servants' indentures, the available evidence, though limited, suggests that neither may have affected the initial colonial auction prices for the servants. This is consistent with the hypothesis that the variable dimensions of the contract were adjusted for the servants' characteristics so as to make the expected auction prices of all servants the same.

The relation between auction prices of servants and the simple cost of passage is also of interest. The modal price of 500 pounds of tobacco observed in the auctions was 11 percent greater than the fare that two free passengers each paid for passage to Virginia on the *Tristam and Jeane* on the same voyage as the servants and 4 percent above the fare quoted elsewhere for passage from England to Maryland in 1638.[10]

[9] This statement neglects one potential element of cost. The marginal cost of delivery, and therefore auction prices, would have included any costs the merchant incurred in recruiting servants in England. These could clearly vary across individual servants, producing differences in expected auction prices. These costs could have been correlated with individual productivity, as in some instances skilled servants were given lump-sum payments at the time of binding. For some evidence of this practice, see the analysis of the Middlesex sample in the Appendix.

[10] The two free passengers each paid 450 pounds of tobacco (London, Public Record Office 1637). For the 1637 Chesapeake farm price of tobacco, see Menard (1975, p. 475). The fare quotation, of £6 sterling per person, is in Maryland Historical Society (1889, p. 206).

That in these cases the typical auction prices tended to be higher than the fare may have been due to the cost of recruiting servants, or to the existence of a premium received by merchants for bearing the risk of servant mortality on the ocean voyage.

Related evidence on the nature of the labor-market equilibrium in the colonies which induced flows of labor from England is available from surviving valuations of indentured servants recorded in Maryland for probate courts. The mean price of 28 male servants with 4 remaining years of servitude recorded on Maryland's lower western shore during 1704–57 was £8.95, with a median of £9 and a mode of £10; while the mean price of 19 females with 4 years remaining was £7.75, with a median of £8 and a mode of £10.[11] The typical marginal cost to an English merchant of delivering servants to the colonies cannot be estimated precisely but probably fell within a range bounded at the lower end by £5, the usual cost of passage, and at the upper end by the £10 estimated by Abbot Emerson Smith as the maximum expense of delivery.[12] The evidence of the probate valuations—although again limited in quantity—suggests that the central tendency of the colonial price of servants did fall within this range and, therefore, offers additional support for the hypothesis that the price of a servant's indenture in the colonies was equal to the marginal cost of delivering labor there from England.

A number of other general considerations relating to the adjustment of the bargain deserve mention. One is the role of mortality. The smaller the probability of a servant's serving a given year of his term, the lower his expected net earnings and, ceteris paribus, the less favorable the terms of the contract he would be able to sign. This is true whether the mortality in question is that during the Atlantic crossing, when the merchant bore the risk, or that after arrival in the colonies, when the planter had purchased the contract and assumed the risk.[13]

A second factor with a similar effect was the possibility of a servant failing to serve out his term for a reason other than death, principally

[11] I am grateful to the St. Mary's City Commission, Annapolis, Maryland, for transcriptions of the probate price quotations, from Maryland Hall of Records. Most of the quotations are from the early decades of the period; on devaluation of Maryland currency and its reflection in probate valuations, see Main (1972, pp. 14–18) and McCusker (1978, pp. 189–204).

[12] Smith 1947, p. 37. A precise analysis would include a positive premium in the probate valuations, for these servants had normally been "seasoned," i.e., had spent a year in the new colonial disease environment, and consequently had a longer life expectancy than the new arrivals, ceteris paribus.

[13] Mortality rates differed among colonies, and the effects of this on the length of indenture will be discussed below. The assumption here will be that mortality rates among servants both during passage and in the colonies were not systematically related to individual productivity.

running away. All the colonies enacted legislation intended to discourage servants from running away; of these some were preventive measures and others punitive. While some colonies provided for corporal—and even capital—punishment for runaways, the most common penalty was extension of the servant's contract by some multiple of the time he was absent.[14] One of the provisions of the contract, the servant's freedom dues, constituted a nonvested pension and, therefore, also acted to discourage servants from running away. While the form and worth of the dues varied across colonies and over time, they were often of substantial value and could have constituted a significant deterrent to servants who considered escaping from their masters.[15]

Servants were not allowed to marry during their terms. Since, by English practice, the expense of raising the illegitimate children of servants fell on the county, colonial legislation provided that the father should be discovered by oath of the mother and that he should reimburse the county for the expense of raising the child until it could be bound out to work. Since servants could not normally pay this sum, either the master would pay it and the servant's term would be extended, or the servant would be bound over to the county for an additional term after the conclusion of his normal term, to be sold for the necessary amount. The mother's term was also extended to reimburse the master for her lost working time.

III. Empirical Results

The principal sources of quantitative evidence bearing on this market for human capital are servant contracts recorded and held in English

[14] For references to laws relating to runaways, see Semmes (1938, pp. 116–18) and Smith (1947, pp. 264–70). The strong terms of this colonial legislation were an important element in the indenture system's success, and some authors have argued that changes in the legal provisions for the enforcement of contracts were central to its decline in the early nineteenth century; see, e.g., Geiser 1901, p. 42. Similarly, the lack of success of the attempt to revive a contract labor system during 1864–85 has been attributed in part to the high cost of enforcing contracts through civil action; see Erickson 1957, pp. 46–48.

[15] On freedom dues, see Smith (1947, pp. 238–41) and Heavner (1976, pp. 50–51). For a discussion of an analogous provision, nonvested pensions as a firm's insurance against quits, see Becker (1975, p. 34). The nature of freedom dues has some implications for other dimensions of the servant contract. The dues were specified by colonial law and were equal for all servants in a given colony. Therefore, the discounted cost of the dues to the planter at the time he purchased a contract varied inversely with the length of the contract. One effect of the fixed nominal value of the freedom dues may, therefore, have been to reduce the amount of variation in the length and other dimensions of the contracts, since the variations in the latter were intended to equalize the net present values of all contracts. This point should not obscure another basic effect of the existence of freedom dues, for, ceteris paribus, they raised the cost of servants and therefore tended to lengthen the term of indenture.

courts. These were made both to protect servants from kidnapping and to protect merchants from false charges of kidnapping. The two largest known surviving collections will be analyzed here; the earlier set, recorded in Middlesex during 1683–84, covers a total of 812 individuals, while the later, recorded in London during 1718–59, covers 3,187 servants.

The principal variable analyzed here will be the length of the indenture. Furthermore, this analysis will be done only for the minors in both samples; these comprise 22 percent of the servants with known ages in the earlier and 67 percent in the later sample. The selection of the variable to be analyzed and of the minors follows from consideration of the evidence of the contracts. It is clear that 4 years was the normal duration of an adult's indenture in both samples, yet for a combination of legal and clerical reasons the full set of conditions of adults' servitude do not appear to have been recorded in either set of indentures. For minors, the full conditions do appear to have been recorded, and the duration of the contract appears to have been the chief variable dimension of the contracts. Both variations in freedom dues and restrictions on the servants' occupations in the colonies were rare. Cash payments made to adults were not generally recorded in the later sample, but the contracts of the minors, on which they do appear to be recorded conscientiously, show that fewer than 6 percent received cash payments. Some cash payments are recorded for both minors and adults in the earlier sample, and although it is uncertain whether payments were recorded in all cases in which they were made, less than 7 percent of the contracts contain such entries. A notable feature of the payments recorded in both samples is that virtually all—95 percent in the earlier and 97 percent in the later sample—were made to servants bound for 4 years. Insofar as payments were made (or promised) and recorded, analysis of the characteristics of servants who received them tends to reinforce the results obtained from the analysis of the length of indenture.[16] The evidence of both samples, therefore, indicates that for minors the greatest variation in the conditions of indenture occurred in the duration of the term of servitude, while for adults this was not the case, as 4 years was both the standard term for adults and normally the minimum term assigned.[17] For adults, cash payments were appar-

[16] See the Appendix.

[17] Terms of less than 4 years do occur, but they appear to have been rare after the midseventeenth century; thus, they account for only 0.5 percent of all indentures of known length in the Middlesex registrations of 1683–84 and 1.2 percent of those in the London registrations of 1718–59. The reasons for this are not known. Four years may have been the term required at most times and places for the average adult to repay the cost of passage out of his net earnings, but it is unclear why shorter terms were not more often given to highly skilled servants like the accountant James Corss, whom

ently substituted for reductions in the length of servitude below 4 years.

Earlier analysis suggested that a servant's term of indenture would be inversely related to the market valuation of his stock of human capital. This index of the servant's human capital can be related to a number of observable characteristics potentially relevant to the determination of the present value of that stock. When this is done by multiple-regression analysis, the estimating equation differs from the common hedonic method only in the use of an index for price. The estimated coefficients of the independent variables age, sex, literacy, and occupation can be interpreted as the marginal prices paid for servants' characteristics in units of the index, while those of destinations represent compensating differentials among regions.[18]

Table 1 shows a number of the basic relationships underlying the market valuation of the servants. In both samples, the length of indenture was negatively related to both age and skill: With other things equal, servants with skilled occupations and those able to sign received shorter terms. Women received shorter terms than men, ceteris paribus, and servants bound for the West Indies received shorter terms than those bound for the North American mainland.

A comparison of the estimated coefficients of the sex variable across samples indicates that, on average, women received considerably larger reductions in their terms in the 1680s than in the eighteenth century. This decline in the premium for females is not surprising in view of the generally declining colonial sex ratios during this period, for while women were preferred for some kinds of household work and some types of farming, their increasing relative availability in most colonies would be expected to lower the size of their wage differentials.[19]

Table 2 provides a more detailed analysis of the length of indenture for the later sample, allowing separate age profiles of length of indenture by sex and skill. It reveals that there was a tendency for women to receive indentures from 5 to 15 percent shorter than those of men through the age of 17, while for servants aged 18–20 there was no

Walter Tullideph sent to the manager of his plantation in Antigua in 1759 with a note stating that he "hath bound himself to serve me four years agreeable to the Laws of Antigua, but as he is 22 years of Age, he thought it hard to serve so long and for that reason, I have given him a Certificate that he is to be absolved from the last year's Service" (Tullideph 1759, vol. 3). That planters preferred to substitute salaries for reductions of the term below 4 years suggests the possibility that fixed costs of hiring and/or a desire to capture the returns from a servant's general training in the colony may have been important considerations.

[18] On the interpretation of coefficients in hedonic price indexes, see Rosen (1974, pp 34–35).

[19] The difference between samples in the sex coefficients in table 1 is significant at the .01 level. On declining colonial sex ratios, see Wells (1975, pp. 156, 219, 244).

TABLE 1

ESTIMATED REGRESSION COEFFICIENTS, MIDDLESEX AND LONDON SAMPLES

	MIDDLESEX, 1683–84		LONDON, 1718–59	
INDEPENDENT VARIABLE	Estimated Coefficient	Standard Error	Estimated Coefficient	Standard Error
Age (years):[a]				
Total sample:				
Less than 15	2.655	.385	2.749	.134
15	2.201	.400	2.147	.080
16	1.457	.304	1.304	.068
17	.893	.367	.728	.062
18	.174	.270	.331	.055
19	.738	.306	.169	.050
Sex[b]	−1.484	.207	−.195	.073
Literacy[c]	−.575	.217	−.082	.037
Date[d]	...	..	−.0060	.0023
Trade[e]	−.727	.445	...	...
Farmer[f]	...	..	−.313	.074
Laborer	...	.	−.146	.079
Services[g]	...	...	−.348	.066
Metal and construction[h]	...	...	−.320	.067
Clothing and textiles[i]	...	...	−.313	.060
Antigua[j]	−.227	.812	−.403	.110
Barbados	−.553	.274	−.176	.154
Jamaica	−.398	.462	−.233	.060
Other West Indies[k]	−.401	1.094	−.479	.088
Maryland	.203	.209	.306	.059
Virginia	...	.	.127	.073
Other mainland[l]	−.389	.673	.050	.116
Constant	5.227	..	4.665	...
R^2	.555	..	.539	..
F	12.87	...	112.82	
n	171	..	2,049	..

SOURCE.—Data used are all from records of minors (age less than 21): Middlesex, 1683–84: London, Greater London Record Office, Nicholson (1965), Wareing (1976); London, 1718–59: London, Corporation of London Records Office, Kaminkow and Kaminkow (1964), Galenson (1977a).

NOTE.—Dependent variable = number of years indentured, method of estimation used is ordinary least squares in tables 1 and 2.

[a] For age variable, indicated age = 1, zero class = age 20.

[b] Male = 0, female = 1.

[c] Marked = 0, signed = 1.

[d] Date entered as final two digits of year of registration.

[e] Trade = 0 for laborers and no recorded occupations, trade = 1 for all other men's occupations.

[f] For all occupational variables, indicated occupation(s) = 1, zero class = no recorded occupation. "Farmer" includes husbandman, plowman, etc.

[g] Includes accountant, barber, surgeon, etc.

[h] Includes blacksmith, carpenter, cooper, mason, etc.

[i] Includes clothier, tailor, weaver, etc.

[j] For all destination variables, for Middlesex sample, zero class = Virginia; for London sample, zero class = Pennsylvania.

[k] Includes Nevis, St. Christopher, etc.

[l] Includes Carolina, New York, etc.

TABLE 2

ESTIMATED REGRESSION COEFFICIENTS, LONDON SAMPLE, 1718–59

Independent Variable	Estimated Coefficient	Standard Error
Age (years):		
Total sample:		
Less than 15	2.976	.144
15	2.378	.092
16	1.542	.084
17	.959	.084
18	.473	.079
19	.260	.083
Women, age:[a]		
Less than 15	−1.034	.460
15	−.472	.390
16	−.969	.228
17	−.302	.161
18	.041	.159
19	.090	.135
20	.198	.156
Trade, age:[b]		
15	−1.496	.392
16	−.884	.206
17	−.502	.117
18	−.275	.085
19	−.224	.073
20	−.103	.076
Literacy	−.076	.036
Date	−.0093	.0024
Antigua[c]	−.260	.235
Barbados	−.005	.268
Jamaica	−.084	.209
Other West Indies	−.363	.221
Maryland	.194	.063
Other mainland	−.039	.072
February[d]	−.100	.092
March	.171	.113
April	.156	.119
May	−.095	.164
June	−.431	.132
July	−.196	.113
August	−.486	.098
September	−.223	.096
October	−.400	.101
November	−.225	.100
December	.014	.097
Sugar[e]	−.0162	.0074
Constant	4.830	. . .
R^2	.566	.
F	52.06	. . .
n	2,044	. .

SOURCE.—London, Corporation of London Records Office, Kaminkow and Kaminkow (1964), Galenson (1977*a*).

NOTES.—See table 1 for variables not defined here.

[a] Female age interactions, indicated variable = 1 for women of given age.

[b] Trade age interactions, indicated variable = 1 for men of given age who recorded a trade (as defined in table 1).

[c] For destinations, zero class = Virginia.

[d] For months, zero class = January. Separate interaction terms between West Indian destination and month of registration were included in the equation, but their coefficients were generally small in value and are not reported.

[e] Sugar = average annual price of muscovado sugar in London, in shillings per hundredweight, lagged 1 year, for servants bound for West Indies (Sheridan [1974, pp. 496–97], with linear interpolation for 1717–20 and 1727).

difference in the length of term by sex. In a suggestive parallel result, Robert Fogel and Stanley Engerman found that, excluding the value of childbearing, the net earnings of female slaves were greater than those of men prior to the age of 18, apparently due to the more rapid physical maturation of women.[20] The sex differentials in the terms of young indentured servants might have resulted from the same source.

The results of table 1 indicate that premia for skills were reflected in the length of servants' indentures. The more detailed specification of the occupations presented in table 1 for the London sample indicates that the marginal premium paid for servants in each of four occupational categories—farmers, services, metal and construction crafts, and clothing and textile trades—was virtually the same.[21] Unskilled laborers received terms longer than servants with skilled occupations but shorter than those with no recorded occupations;[22] the latter result may indicate that some premium was paid for the laborers' work experience.

Economists have devoted considerable attention to the analysis of the relationship between productivity and age and have accumulated much evidence on the association between age and wages in recent periods. Less is known of the nature of this relation in past times. It is, therefore, of some interest to consider in more detail the implications of the estimated relationships between age and length of indenture for the age-earnings profiles of servants.[23] Table 3 presents estimates of the relative annual net earnings of servants by age for unskilled and skilled men. The calculations are based on the assumption that the expected colonial sale price of each individual's contract was equal to the constant marginal cost of delivering servants to the colonies.[24] The relationship between age and net earnings is assumed to have been linear, making average net productivity during the term equal to net productivity at the term's midpoint. The estimates of average net

[20] Fogel and Engerman 1974, p. 77, see also Metzer 1975, pp. 136–37. Interestingly, evidence on the hourly earnings of North Carolina cotton mill employees in 1907 indicates that average female earnings were above those of males through the age of 15, equal at 16, and below male earnings thereafter (Wright 1980, p. 6).

[21] None of the four coefficients is significantly different from any of the other three at the .10 level. On the value of skilled servants in the colonies, see, e.g., Martin (1761, vol. 4, fol. 97, verso); Jeaffreson (1878, 1:186); and Galenson (1979*a*, pp. 314–19).

[22] The coefficient of laborer for the London sample in table 1 is significantly different from that of farmer at the .10 level for a one-tailed *t*-test, from those of metal-construction and clothing-textiles at .05, and from that of services at .025.

[23] It might be noted that the erratic behavior of the coefficients of higher ages in table 1 for the Middlesex sample may have been due to the falsification of the ages of some servants. This may have resulted from the legal requirements under which the registrations were made; for discussion and evidence see Galenson (1979*a*, appendix to chap. 3).

[24] Possible seasonal variation in delivery costs has been controlled for in the equation reported in table 2

productivity are derived from the following formulation of the mean present value of the contracts of servants of age j at the beginning of the term:

$$PV_j = \sum_{i=1}^{n} \frac{NP_i - w_i}{(1 + r)^i},$$

where NP = expected mean annual net productivity during the term; n = mean length of term for servants in each entering age group; w = mean annual wage payments made to servants during the term; r = discount rate.

To solve for the value of NP for each entering cohort, the mean present value of the contracts was set equal to £10, an estimate of the marginal cost of delivering servants to the colonies. The mean age of each group at the time of binding was taken as the recorded age plus one-half year to allow for the rounding of age upon registration. The mean length of term by age was derived from the coefficients of table 2, while the mean annual wage payments were taken directly from the indenture contracts.[25] The estimates were made with a discount rate of 10 percent.

The estimates of table 3, which indicate that the net-earnings profile of skilled servants was steeper than that of the unskilled, are consistent with the normal positive relationship between the steepness of age-earnings profiles and training. The ages at which the servants considered here were bound, between 15 and 20, were prime ones for training in a wide variety of skilled crafts in preindustrial England, through either apprenticeship or less formal arrangements. It is, therefore, not surprising that the net productivity of those in skilled trades rose rapidly during this period of the life cycle.

Skilled servants received a considerable premium: Even at age 15, a

[25] No wages were recorded for unskilled servants. The average wage payments received by the skilled servants in the London sample by age were as in the table below. The present value of freedom dues at the time of binding varied across colonies and over time, according to differences and changes in legislation and with changes in the value of colonial currencies and commodities. No explicit allowance has been made for the dues in the calculation because of the difficulty of estimating their typical value; inclusion of the effect of the lump-sum payment would lower the estimates of the net annual earnings of servants without changing their relative values by age.

Age (Years)	Mean Annual Wage (£)
15	.0
16	.40
17	.79
18	1.45
19	1.52
20	3.02

TABLE 3

ESTIMATED RELATIVE NET ANNUAL EARNINGS OF SERVANTS BY AGE

Unskilled			Skilled		
Age	Mean Net Annual Earnings (£)	Relative Net Earnings (Age 22.9 = 1.00)	Age	Mean Net Annual Earnings (£)	Relative Net Earnings (Age 22.9 = 1.00)
19.1	2.01	.742	18.4	2.39	.413
19.7	2.20	.812	19.2	2.86	.495
20.4	2.36	.871	20.1	3.32	.574
21.2	2.52	.930	21.0	4.08	.706
22.0	2.60	.959	21.9	4.22	.730
22.9	2.71	1.000	22.9	5.78	1.000

NOTE.—Calculated from table 2 and n. 24. See text for procedure. The unskilled profile is calculated from the basic age profile of table 2, that of the skilled from the basic age profile combined with the skilled ("trade") coefficients.

skilled servant received a term 21 percent shorter than his unskilled counterpart. The existence of a sizable premium at such an early age could have been due in part to differences in the relative average work experience of the skilled and unskilled. Thus, possession of a skilled trade at any age implied prior work experience. The age of entry into the labor force for those men registered without occupations cannot be determined, but it is possible that the typical age of entry for the unskilled into employment was that at which English boys normally left home to live in service, roughly 15. If this were the case, work experience and acquired on-the-job training might have accounted for a significant portion of the premium for skilled workers. This would particularly be true for the younger servants, as the relative level of work experience of a skilled to an unskilled worker would be greatest at the lower ages observed here and would decline with age thereafter. That the ratio of skilled to unskilled net earnings increased with age would appear to be strong evidence of the presence of formal training for those in the skilled group.

An interesting feature of the relative age–net-earnings profile of the unskilled shown in table 3 is its close resemblance to those profiles obtained by Fogel and Engerman for unskilled male slaves in the southern United States during 1790–1860 (1972, charts 3 and 4). In view of the considerable differences among these samples with respect to such variables as location and crops cultivated, the similarity of the shapes of the profiles might suggest the importance of physiological factors, particularly the rate of physical maturation, in determining the change, with age, in the productivity of unskilled workers

under conditions of plantation agriculture in the eighteenth and nineteenth centuries.

The relatively small premium paid for the ability to sign in the later sample may have been due both to the abundance of literate servants and to high literacy rates in the colonies. The decline in the size of the reduction of the term due to the ability to sign between the dates of the two samples may have resulted in part from a considerable increase in literacy among the servants, as only 35 percent signed in the earlier sample compared with 67 percent in the later one.[26]

In both samples, with other characteristics constant, servants bound for the West Indies tended to receive shorter terms than those bound for mainland colonies. That servants who immigrated to the West Indies received shorter terms to compensate them for their choice is consistent with the fact that, while both working conditions for servants and economic opportunities for freedmen were known to be poor in the islands after the introduction of large-scale sugar cultivation in the second half of the seventeenth century—with its attendant slave gangs and consolidation of small farms into large estates—the mainland long continued to be considered a land of opportunity for poor immigrants, where freed servants could hope to own land and become prosperous members of society. A persistent theme of West Indian complaints appeared in a 1675 petition sent to the king of England by the Council and Assembly of Barbados: "In former tymes Wee were plentifully furnished with Christian [i.e., white] servants from England . . . but now Wee can gett few English, having noe Lands to give them at the end of their tyme, which formerly was their main allurement." The higher mortality rates of the West Indian colonies decreased servants' expected productivity and made them reluctant to go to the region. However, in conjunction with the high productivity of labor in sugar production, those rates acted to raise the demand for new flows of replacement immigrant labor. That servants bound for the West Indies received terms shorter than those bound for the mainland, in spite of the higher mortality rates in the islands, implies that the marginal productivity of labor was higher in the West Indies than in the mainland colonies.[27]

Both the lower estimated intercept for the later sample and the estimated negative time trend of the later sample indicate a secular decline in the length of indenture. The direction of change is consis-

[26] On the relationship between ability to sign and other aspects of literacy in this period, see Schofield (1968, pp. 311–25). The difference between the coefficients of literacy in the two samples is significant at .10 for a two-tailed t-test. On the servants' ability to sign, see Galenson (1979*b*).

[27] London, Public Record Office 1675. On relative mortality rates in the West Indian and mainland colonies, see Wells (1975, pp. 280–82).

tent with a number of long-term trends, including rising reservation wages of servants due to rising real wages in England between the mid-1680s and the middle of the eighteenth century, falling real shipping costs, and declining colonial mortality rates which could have produced a secular increase in the colonial demand for labor.[28] The negative estimated trend further suggests the presence of a secular increase in real wages in the colonies during the period spanned by these two samples.

A consistent seasonal pattern in the length of indenture appears in the results of table 2, as servants bound for mainland destinations whose indentures were signed between June and November received sizable reductions in their terms relative to servants bound in winter and spring. Most of these servants were bound for Maryland or Virginia, and the observed pattern could have been due to the effect of the seasonality of tobacco production on the costs involved in supplying servant labor to the Chesapeake. The shipping patterns which resulted from the timing of the harvests dictated that the amount of backhaul space for servants was greatest in summer and early fall. Since servants were provided with food and lodging from the time they signed their contracts, the cost of delivering a servant to the colonies may have declined in peak shipping seasons because the more frequent departures of ships reduced the average waiting time in port between binding and sailing. The lower costs of the peak seasons could, therefore, have resulted in shorter indentures for servants bound in peak seasons than for those bound in slack shipping months.[29]

Another potential source of variation in the length of indenture was annual changes in the colonial demand for labor. The results presented in table 2 indicate that the lagged annual average price of muscovado sugar in London had a significant and negative effect on the length of indentures of servants bound for the West Indies during 1718–59; the estimated effect of a change in the price of sugar from its minimum to its maximum in the period, with other things equal,

[28] On English wages, see Gilboy (1934, pp 219–25) and Phelps Brown and Hopkins (1956, pp. 302–13). On changes in colonial mortality rates, see, e.g., Menard (1977*a*, pp. 99–100) The organization and quality of information in the market for contracts may have improved during the period spanned by the two samples. It is suggestive that the coefficient of variation of the term of indenture among men of a given age was considerably lower in the later than the earlier sample for six of the eight age groups of minors above the age of 12. On the relation of wage dispersion to information, see Stigler (1962).

[29] On the seasonality of tobacco production and shipping, see Bullock (1649, p. 46); Alsop (1666, p. 51); and Bruce (1907, vol. 1, pp. 622–24). On the costs of maintaining servants between binding and sailing, see Smith (1947, pp. 36–37, 59–65) and Scottish Record Office (GD 23/6/98, nos 4, 14, 18). For factors relevant to seasonality in the colonial demand for labor, see Mullin (1972, p. 15) and Morgan (1975, p. 158).

was a reduction of about 5 months in the term of indenture. The sign of this effect would be the one predicted if, as appears to have been the case, high sugar prices normally resulted from high levels of demand for sugar rather than reductions in supply, for high sugar prices would then have tended to produce high demand for labor and, ceteris paribus, to shorten terms.[30]

IV. Conclusions

The price paid for human capital in the colonial American market for indentured servants varied systematically with respect to factors which influenced servants' productivity, as economic theory predicts. All servants incurred debts of similar value in immigrating to the colonies and sold claims on their future labor, in the form of indentures, to repay these debts. Characteristics which raised the expected productivity of servants in the American colonies raised the market valuation of their human capital and, therefore, shortened the term for which the servant was bound. Thus, servants with skilled trades and those able to sign served shorter terms than the unskilled and illiterate of similar age and sex. Women were found to have received shorter terms than men until the age of 18, perhaps due to their earlier maturation. The results also indicated that servants bound for the West Indies received shorter terms in compensation for their undesirable destinations. Servants bound during peak shipping seasons were found to have received reduced terms, perhaps due to the shorter average waiting time prior to departure in those months when backhaul cargo space was most abundant. Finally, the length of indenture of servants bound for the West Indies was found to have varied inversely with the price of sugar, suggesting that increases in the colonial demand for labor shortened the term of servitude.

Among the issues which need further investigation are the precise reasons for the observed shapes of the age-earnings profiles of servants. Additional research, including the collection of evidence on wage rates in the colonies by age and skill, may serve to distinguish

[30] Similar analysis of the contracts of servants bound for the Chesapeake during 1718–40 indicates that the lagged annual farm price of Maryland tobacco had no significant effect on the length of indenture; in a regression equation which included the variables of table 2, the estimated coefficient of an interaction term between the price of tobacco, lagged 1 year, and Chesapeake destination was insignificant. Due to a greater continuing reliance on production of a single staple in the West Indies in the eighteenth century, the price of sugar may serve as a better index of the West Indian demand for labor than does the price of tobacco for the Chesapeake. On diversification of agricultural production in the Chesapeake, see Clemens (1974, pp. 100–148) and Stiverson (1977, pp. 65–103).

and isolate the effects of such contributory factors as physical maturation and investment in human capital in producing the age-earnings relationships which lay behind the market valuation of indentured human capital. What the present research has indicated is that the application of economic analysis to quantitative evidence generated by the system of indentured servitude can provide information on the way in which the market once explicitly evaluated stocks of human capital and, in so doing, can yield new insights into the operation of labor markets in early America.

Appendix

For the Middlesex sample of 1683–84, a regression equation was estimated with the amount of the cash payment (in shillings, sterling) made to a servant as the dependent variable, with the same independent variables used in the analysis of the length of indenture. The sample used was that of all servants, minors and adults, whose contracts contained all the necessary information. The results are as shown in table A1.

The hypothesis that all the coefficients are simultaneously equal to zero can be rejected at the .01 level, but the proportion of the variance explained is small, as might be expected in view of the rarity of the payments. Only one coefficient is significant at the .01 level: Possession of a skilled trade, which table 1 shows to have reduced the term of an indenture, significantly raised the expected cash payment made to a servant.

As noted in the text, cash payments appear to have been recorded only for minors in the London sample of 1718–59. The form of the payments recorded differs from that of the earlier sample: Whereas the cash payments to servants in the Middlesex contracts of 1683–84 appear to have been simple lump-sum payments made at the time of binding, those contracted for in London in the eighteenth-century sample were generally salaries to be paid

TABLE A1

ESTIMATED REGRESSION COEFFICIENTS, MIDDLESEX SAMPLE, 1683–84

Independent Variable	Estimated Coefficient	Standard Error
Age (years)	.011	.038
Sex[a]	−.485	.456
Literacy[b]	−.443	.363
Trade[c]	1.466	.399
West Indies[d]	.088	.354
Constant	.323	...
R^2	.026	...
F	4.03	..
n	760	...

SOURCE.—London Greater London Record Office (Middlesex Section)
[a]Male = 0, female = 1
[b]Marked = 0, signed = 1.
[c]Laborer or no occupation = 0; all other occupations = 1
[d]West Indian destinations = 1, mainland = 0

TABLE A2

ESTIMATED REGRESSION COEFFICIENTS, LONDON SAMPLE, 1718–59

Independent Variable	Estimated Coefficient	Standard Error
Age (years)	.144	.054
Sex	.306	.329
Literacy	.428	.167
Trade	1.397	.190
West Indies	1.291	.165
Constant	−3.141	. .
R^2	.094	. . .
F	42.20	. . .
n	2,046	. . .

SOURCE.—London, Corporation of London Records Office.
NOTE.—All variables defined as in table A1.

annually in local currency. A regression was estimated with the amount of the annual salary (in pounds local currency) as the dependent variable, with the same independent variables used above. (The dependent variable is the unadjusted value of the colonial currency. Although most of the colonies' currencies were devalued relative to sterling, the differences in currency values across the principal American colonies in most of this period were small. See McCusker 1978.) The sample used was that of all minors (age less than 21) whose contracts contained all the necessary information. The results were as shown in table A2.

The relationship is again statistically significant, and the proportion of the variance explained is again low. Age, literacy, possession of a skilled trade, and West Indian destinations are all significant at .01 and positively related to the servant's salary; in table 2 it is shown that all are significantly and negatively associated with the length of indenture.

The results of the analysis of servants' cash payments and salaries in both samples are consistent with the hypothesis that these were positively related to the servants' expected earnings in the colonies. These results therefore reinforce the analysis presented of the length of indenture.

References

Alexander, Edward P., ed. *The Journal of John Fontaine*. Charlottesville: Univ. Press Virginia, 1972.

Alsop, George. *A Character of the Province of Mary-Land*. London: Peter Dring, 1666.

Annapolis. Maryland Hall of Records. Maryland Probate Records.

Becker, Gary S. *Human Capital*. 2d ed. New York: Columbia Univ. Press (for Nat. Bur. Econ. Res.), 1975.

Bruce, Phillip Alexander. *Economic History of Virginia in the Seventeenth Century*. New York: Macmillan, 1907.

Bullock, William. *Virginia Impartially Examined, and Left to Publick View, to Be Considered by All Iudicious and Honest Men*. London: John Hammond, 1649.

Clemens, Paul G. E. "From Tobacco to Grain: Economic Development on

Maryland's Eastern Shore, 1660–1750." Ph.D. dissertation, Univ. Wisconsin, 1974.

Dunn, Richard S. *Sugar and Slaves: The Rise of the Planter Class in the English West Indies, 1624–1713.* New York: Norton, 1973.

Edinburgh. Scottish Record Office. Bught Papers. GD 23/6/98.

Erickson, Charlotte. *American Industry and the European Immigrant, 1860–1885.* Cambridge, Mass.: Harvard Univ. Press, 1957.

Fogel, Robert William, and Engerman, Stanley L. "The Market Evaluation of Human Capital: The Case of Slavery." Paper presented at the Cliometrics Conference, Madison, Wis., 1972.

———. *Time on the Cross: The Economics of American Negro Slavery.* Boston: Little, Brown, 1974.

Galenson, David W. "Agreements to Serve in America and the West Indies, 1727–31." *Genealogists' Magazine* 19 (June 1977): 40–44. (*a*).

———. "Immigration and the Colonial Labor System: An Analysis of the Length of Indenture." *Explorations Econ. Hist.* 14 (October 1977): 360–77. (*b*).

———. "The Indenture System and the Colonial Labor Market: An Economic History of White Servitude in British America." Ph.D. dissertation, Harvard Univ., 1979. (*a*).

———. "Literacy and the Social Origins of Some Early Americans." *Historical J.* 22 (March 1979): 75–91. (*b*).

Geiser, Karl Frederick. *Redemptioners and Indentured Servants in the Colony and Commonwealth of Pennsylvania.* New Haven, Conn.: Tuttle, Morehouse & Taylor, 1901.

Gilboy, Elizabeth W. *Wages in Eighteenth Century England.* Cambridge, Mass.: Harvard Univ. Press, 1934.

Gray, Lewis Cecil. *History of Agriculture in the Southern United States to 1860.* Gloucester, Mass.: Peter Smith, 1958.

Greene, Evarts B., and Harrington, Virginia D. *American Population before the Federal Census of 1790.* New York: Columbia Univ. Press, 1932.

Greene, Lorenzo Johnston. *The Negro in Colonial New England, 1620–1776.* New York: Columbia Univ. Press, 1942.

Heavner, Robert O. "Economic Aspects of Indentured Servitude in Colonial Pennsylvania." Ph.D. dissertation, Stanford Univ., 1976.

Jeaffreson, John Cordy, ed. *A Young Squire of the Seventeenth Century. From the Papers (A.D. 1676–1686) of Christopher Jeaffreson.* London: Hurst & Blackett, 1878.

Kaminkow, Jack, and Kaminkow, Marion. *A List of Emigrants from England to America, 1718–1759.* Baltimore: Magna Charta, 1964.

Kingsbury, Susan M., ed. *The Records of the Virginia Company of London.* Washington: Government Printing Office, 1906.

Littleton, Edward. *The Groans of the Plantations.* London: M. Clark, 1689.

London. Corporation of London Records Office. "Memoranda of Agreements to Serve in America and the West Indies."

London. Greater London Record Office (Middlesex Section). "Plantation Indentures." MR/E.

London. Public Record Office. Colonial Office, class 1, piece 35, f. 237v. "Petition of the Council and Assembly of Barbados to the King." 1675.

———. High Court of the Admiralty, class 20, piece 636. "Ledger for Goodes Sould in ye Barbadoes . . . Sent by ye Shipp Abraham 1636. . . ." 1636.

———. High Court of the Admiralty, class 20, piece 635. "Booke of Accompte

for the Shippe Called ye Tristam and Jeane of London wch Came from Virginia Anno Dm 1637." 1637.

McCusker, John J. *Money and Exchange in Europe and America, 1600–1775: A Handbook.* Chapel Hill: Univ. North Carolina Press, 1978.

Main, Gloria Lund. "Personal Wealth in Colonial America: Explorations in the Use of Probate Records from Maryland and Massachusetts, 1650–1720." Ph.D. dissertation, Columbia Univ., 1972.

Martin, Samuel. "Letter Book of Samuel Martin Senior." British Museum, Martin Papers, Add. MSS 41,349, 1761.

Maryland Historical Society. *The Calvert Papers.* Baltimore: Maryland Hist. Soc., 1889.

Menard, Russell R. "Economy and Society in Early Colonial Maryland." Ph.D. dissertation, Univ. Iowa, 1975.

———. "Immigrants and Their Increase: The Process of Population Growth in Early Colonial Maryland." In *Law, Society, and Politics in Early Maryland,* edited by Aubrey C. Land, Lois G. Carr, and Edward C. Papenfuse. Baltimore. Johns Hopkins Univ. Press, 1977. (*a*)

———. "From Servants to Slaves: The Transformation of the Chesapeake Labor System." *Southern Studies* 16 (Winter 1977): 355–90. (*b*)

Metzer, Jacob. "Rational Management, Modern Business Practices, and Economies of Scale in the Ante-bellum Southern Plantations." *Explorations Econ. Hist.* 12 (April 1975): 123–50.

Middleton, Arthur Pierce. *Tobacco Coast: A Maritime History of Chesapeake Bay in the Colonial Era.* Newport News, Va.: Mariners' Museum, 1953.

Morgan, Edmund S. *American Slavery, American Freedom: The Ordeal of Colonial Virginia.* New York: Norton, 1975.

Mullin, Gerald W. *Flight and Rebellion: Slave Resistance in Eighteenth-Century Virginia.* London: Oxford Univ. Press, 1972.

Nicholson, Cregoe D. P. *Some Early Emigrants to America.* Baltimore: Genealogical Pub. Co., 1965.

Pares, Richard. *Merchants and Planters. Econ. Hist. Rev.,* suppl. no. 4. Cambridge: Cambridge Univ. Press, 1960.

Phelps Brown, E. H., and Hopkins, Sheila V. "Seven Centuries of the Price of Consumables, Compared with Builders' Wage-Rates." *Economica* 23 (November 1956). 296–314.

Purchas, Samuel. *Purchas His Pilgrimes.* London: Henrie Featherstone, 1625.

Rosen, Sherwin. "Hedonic Prices and Implicit Markets: Product Differentiation in Pure Competition." *J.P.E.* 82, no. 1 (January/February 1974): 34–55.

Schofield, R. S. "The Measurement of Literacy in Pre-industrial England." In *Literacy in Traditional Societies,* edited by Jack Goody. Cambridge: Cambridge Univ. Press, 1968.

Semmes, Raphael. *Crime and Punishment in Early Maryland.* Baltimore: Johns Hopkins Univ. Press, 1938.

Sheridan, Richard B. *Sugar and Slavery: An Economic History of the British West Indies, 1623–1775.* Barbados: Caribbean Univ. Press, 1974.

Smith, Abbot Emerson. *Colonists in Bondage: White Servitude and Convict Labor in America, 1607–1776.* Chapel Hill: Univ. North Carolina Press, 1947.

Smith, John. *The Generall Historie of Virginia, New-England, and the Summer Isles.* London: Michael Sparkes, 1624.

Stigler, George J. "Information in the Labor Market." *J.P.E.* 70, no. 5, pt. 2 (October 1962): 94–105.

Stiverson, Gregory A. *Poverty in a Land of Plenty: Tenancy in Eighteenth-Century Maryland.* Baltimore: Johns Hopkins Univ. Press, 1977.

Sutherland, Stella H. *Population Distribution in Colonial America.* New York: Columbia Univ. Press, 1936.

Taunton. Somerset Record Office. Helyar Manuscripts.

Tullideph, Walter. "Letter Books of Dr. Walter Tullideph." Unpublished. Dundee, Scotland, 1759.

Wareing, John. "Some Early Emigrants to America, 1683–84: A Supplementary List." *Genealogists' Magazine* 18 (March 1976): 239–46.

Wells, Robert V. *The Population of the British Colonies in America before 1776.* Princeton, N.J.: Princeton Univ. Press, 1975.

Wilson, Samuel. *An Account of the Province of Carolina in America.* London: Francis Smith, 1682.

Wood, Peter H. *Black Majority: Negroes in Colonial South Carolina from 1670 through the Stono Rebellion* New York: Norton, 1975.

Wright, Gavin. "Cheap Labor and Southern Textiles, 1880–1930." Paper presented at the Workshop in Economic History. Univ. Chicago, January 25, 1980.

The Welfare Cost of Capital Income Taxation in a Growing Economy

Christophe Chamley

Yale University

The welfare cost of capital income taxation is analyzed in a general equilibrium framework, where the private sector is represented by a competitive household endowed with perfect foresight and an infinite life. The value of the welfare cost depends essentially on the elasticity of substitution between capital and labor in the production function. Numerical estimates are presented for different values of the parameters of the model. The welfare gain obtained by the abolition of the capital income tax is smaller when the private sector is not endowed with perfect foresight (it is reduced by about 40 percent when expectations are myopic). The allocation efficiency cost of the corporate tax dwarfs the intertemporal welfare cost.

A central issue in the current debate on tax reform is the efficiency cost of a tax on capital income. Such a tax introduces a wedge between the prices of consumption at different dates and distorts the intertemporal allocation of resources.

Previous studies on the efficiency cost of the capital income tax have relied on simplifying assumptions which rule out important consequences of capital taxation. In general, these studies are of two types. The first (Krzyzaniak 1967; Sato 1967; Feldstein 1974*a*, 1974*b*; Friedländer and Vandendorpe 1978) analyzes the capital income tax in the context of a neoclassical growth model with endogenous capital stock and factor prices but does not allow an optimal response of

I am grateful to the Cowles Foundation for financial support. Comments by Laurence Kotlikoff, Dale Jorgenson, Laurence Weiss, John Shoven, and anonymous referees were very helpful. This paper would not have been completed without the collaboration of William Brainard.

Journal of Political Economy, 1981, vol. 89, no. 3]
© 1981 by The University of Chicago 0022-3808/81/8903-0014$01.50

household to changes of taxes. Typically the levels of consumption and labor supply do not depend on future prices and are not consistent with utility maximization. This method leads to results about the long-run incidence of taxation on aggregate variables; the transition to the steady state can be examined by numerical simulations. The second type of study (Levhari and Sheshinski 1972; Feldstein 1978) allows for maximizing behavior but assumes that the factor prices are exogenous and independent of the capital accumulation. The welfare cost of the capital income tax is measured by an application of the Harberger-Hicks-Hotelling formula (Green and Sheshinski 1979) to the intertemporal framework.

The welfare cost of the capital income tax is analyzed here in a stylized intertemporal general equilibrium model (described in the next section). Issues about intra- or intergenerational equity are ignored. Therefore, it is assumed that the private sector can be represented by a household with an infinite life.[1]

This household determines the level of consumption by the maximization of its intertemporal utility function. Future factor prices (wage and interest rates) depend on the accumulation of capital through a neoclassical technology and are known with perfect foresight. The household behaves competitively: Future (endogenous) prices are taken as given.

The welfare cost of the capital income tax is analyzed in the second section. Following common practice, we consider the welfare cost induced by a tax with lump-sum redistribution. Initially, the economy is assumed to be on the balanced growth path where capital income is taxed at a fixed rate (and tax revenues are refunded). At time zero, the tax is abolished (together with the refunds). Thereafter, the economy moves on a dynamic path toward a new steady state. The welfare cost of the capital income tax is equal to the welfare gain obtained by the abolition of the tax, namely, by the difference between the level of utility on the new dynamic path (after the tax reform) and the level of utility on the initial balanced growth path (with the tax in effect).[2] As usual, the welfare cost of the tax is

[1] It would be sufficient to assume that individuals' utilities depend on their consumption in their own finite lifetime and on the welfare of their immediate descendants (for a development of this argument, see Barro [1974]). If individuals do not leave a bequest, the problem of intergenerational equity arises, and the concept of excess burden used here does not apply: Even in the case where no revenue has to be raised, the no-tax solution is not the first best; the level of capital has no optimal property (Diamond 1965). In order to redistribute income among generations, a benevolent social planner would impose a tax or a subsidy on capital income (and possibly on other goods). For an analysis of taxation in a general equilibrium model where individuals maximize a lifetime utility function with no bequest, see Hall (1969), Diamond (1970), Summers (1979), and Chamley (1980*b*).

[2] We could also consider the imposition of a tax on the initial balanced growth path

measured by a wealth equivalent and is of a second order with respect to the tax rate. A second approximation of this excess burden is given, which depends on the parameters of the utility and production functions and on the growth rate. An extension of the Levhari and Sheshinski (1972) result is obtained as a special case. Since the general excess-burden formula is exact only for infinitesimal values of the tax rate, the error of this second-order approximation is analyzed in a numerical example.

The assumption of perfect foresight is relaxed in Section III. Because the dynamic path after the abolition of the tax is no longer optimal, the welfare gain induced by the tax reform is smaller in this case. The case of myopic expectations is an important example of the more general class of expectations which are considered. In the following sections we revert to the assumption of perfect foresight.

The case where the tax rate on capital income is not identical for all sectors of production (an example is found in the corporation tax) is considered in Section IV. The intertemporal welfare cost of the tax is compared with the inefficiency cost due to the misallocation of capital between the different sectors of production.

In Section V, the assumption of a fixed labor supply is relaxed. Since the capital income tax lowers the long-run wage rate, its excess burden depends on the (compensated) elasticity of the labor supply.

This analysis of the capital income tax relies on a stylized model.[3] However, numerous numerical examples show that some of the results obtained are fairly robust. In the conclusion, these results are summarized and related to other studies using more disaggregated models.

I. The Model

There is one good in the economy. This good can be consumed or used as capital in the production process. Total output per efficiency unit of labor, net of capital depreciation, is given by the neoclassical production function, $y = f(k)$, where k is the level of the capital stock per efficiency unit of labor.

The private sector is represented by a household, growing at the rate n, which takes prices as given and maximizes under its budget constraint the utility function,

$$U = \int_0^\infty e^{-\rho t} e^{nt} u(c_t e^{\mu t}) dt, \tag{1}$$

with no taxation. However, the analysis of the abolition (instead of the imposition) of the tax is technically more simple and suits better the problems of tax reform. It has been verified that for small tax rates, the two methods give the same results.

[3] The same method could have been used to analyze the excess burden of the labor income tax or the income tax (Chamley 1980*a*).

where the following notation is used: ρ = pure rate of time preference, μ = rate of growth of labor augmenting technological change, and c_t = consumption per unit of efficient labor.

The function u will be assumed to be of the form $u(c) = c^{1-\sigma}$.[4] The labor supply per capita is fixed and normalized to one at time zero. The representative household is endowed with perfect foresight and behaves competitively, taking the endogenous future prices (wage and interest rates) as given.

Because of the first-order condition in the maximization of the utility function, the dynamic path of the economy satisfies the equation[5]

$$\dot{c}_t = \frac{c_t}{\sigma}(r_t - \rho^*), \tag{2}$$

where r_t is the net rate of return available to the household, and $\rho^* = \rho + \sigma \cdot \mu$.

The capital accumulation is defined by

$$\dot{k}_t = f(k_t) - (n + \mu)k_t - c_t, \tag{3}$$

where k_0 represents the initial capital stock.

The dynamic behavior of the economy is defined by equations (2) and (3) and by the initial values k_0 and c_0 at time zero. The initial value of the capital stock k_0 is given. There is a unique value of c_0, such that its associated dynamic path satisfies the budget constraint of the household (which is equivalent here to the transversality conditions). For this value of c_0, the dynamic path converges to the steady state defined by[6]

$$\rho^* = r^* = f'(k^*), \tag{4}$$

$$c^* = f(k^*) - (n + \mu)k^* \tag{5}$$

(an asterisk will denote a steady-state value).

[4] This assumption is necessary for the existence of a competitive balanced growth path when the intertemporal welfare function is additive and the rate of labor augmenting technological change μ is different from zero. (For an introduction to the literature on optimal growth, see Koopmans [1967].) In a discrete-time formulation one could also use a stationary utility function. The existence and stability of optimal balanced growth paths in this context is studied by Iwai (1972).

[5] The consumption levels at time t and $t + \Delta t$ satisfy the relation,

$$\frac{u'[e^{\mu(t+\Delta t)}c_{t+\Delta t}]}{u'(e^{\mu t}c_t)} = \frac{1 + \rho\Delta t}{1 + r\Delta t}.$$

Up to the first order, this expression is equivalent to

$$\frac{u'(e^{\mu t}c_t) + u''(e^{\mu t}c_t)(c_t\mu\Delta t + \Delta c_t)e^{\mu t}}{u'(e^{\mu t}c_t)} = 1 + (\rho - r)\Delta t.$$

A straightforward manipulation gives $(1/c_t)(\Delta c_t/\Delta t) = (1/\sigma)(r - \rho - \sigma\mu)$. When Δt tends to 0, we obtain eq. (2).

[6] The second-order conditions are derived from the concavity of the utility and the production functions. For an exhaustive treatment, see Arrow and Kurz (1970).

The optimal path also defines a consumption function, giving the level of consumption per unit of labor as a function of the capital-labor ratio at each instant (see fig. 1),

$$c = c(k). \tag{6}$$

In the same way, at a given instant, the level of utility is determined by the integral (1) on the optimal path Γ for an initial value of the capital stock which is equal to k: $U = J(k)$.

We now review a few properties of the optimal path which will be useful in the subsequent sections.

The slope of the consumption function $c'(k^*)$ at the stationary point k^* is obtained by taking the limit of the ratio between the relations (2) and (3) when k tends to k^*; $c'(k^*)$ is equal to the positive root of the equation,

$$x^2 - \lambda x - \gamma = 0, \tag{7}$$

where $\lambda = \rho^* - n - \mu$ and $\gamma = -[c(k^*)f''(k^*)]/\sigma = (1/\sigma\epsilon)(c(k^*)(r^* + \delta)w^*/\{k^*[f(k^*) + \delta k^*]\})$; ϵ is the elasticity of substitution between capital and labor in the *gross* production function when the capital labor ratio is equal to k^*.[7] Using a first-order approximation of the capital accumulation (3) around k^*, the difference between k and its steady-state value k^* decreases asymptotically at a constant rate a (to a first-order approximation):[8]

$$(k_t \dot{-} k^*) = -a(k_t - k^*), \tag{8}$$

where a is the coefficient of adjustment of the economy toward the steady state and is equal to the difference $c'(k^*) - \lambda$; $-a$ is also equal to the negative root of equation (7).

The same asymptotic rule applies to every endogenous variable z_t in the economy which depends only on the capital-labor ratio (as, e.g., the gross factor prices):[9]

$$(z_t \dot{-} z^*) = -a(z_t - z^*), \tag{8a}$$

where the coefficient a is the same as in equation (8).

[7] The parameter γ can also be expressed as a function of quantities which are easily measurable:

$$\gamma = \frac{(1 - \alpha)(r^* + \delta)}{\sigma\epsilon}\left[\frac{r^* + \delta}{\alpha} - (\delta + n + \mu)\right],$$

where α is the share of gross capital income in the production function and δ is the depreciation rate of the capital stock.

[8] This regressive rule may be a very good approximation even if the difference $k - k^*$ is large (see Chamley 1979).

[9] If $z_t = g(k_t)$, to the first order: $(z_t - z^*) = g'(k^*)k_t = -g'(k^*)a(k_t - k^*) = -a(z_t - z^*)$.

II. The Excess Burden of the Capital Income Tax

Consider now a tax on capital income at the constant rate θ, with lump-sum redistribution of its revenues. The analysis of the model described in the previous section applies with a net interest rate r now given by $r = (1 - \theta)f'(k)$. In particular, in the long run, the net rate of return is constant and still equal to ρ^*. In the long run, the capital income tax increases the gross rate of return and lowers the capital stock and aggregate consumption. The levels of consumption and capital per unit of labor in the steady state with taxation, $\bar{c}$ and $\bar{k}$, respectively, are given by

$$\rho^* = (1 - \theta)f'(\bar{k}) \tag{4a}$$

and

$$\bar{c} = f(\bar{k}) - (n + \mu)\bar{k}. \tag{5a}$$

This steady state is represented by point A in figure 1.

We assume that initially the economy is in the long-run equilibrium with taxation (fig. 1, point A). The deadweight loss of the tax is measured by the welfare gain induced by the suppression of the tax.[10]

At time zero, the capital income tax is abolished[11] (together with the lump-sum refund). The net rate of return suddenly increases, and consumption decreases immediately from $\bar{c}$ to c_0 (the optimal value which depends only on the initial capital stock $k_0 = \bar{k}$, as described in Sec. I). After time zero, because of the increased savings,[12] the capital stock increases. The economy moves toward the new steady state E on the segment BE of the path Γ defined by the dynamic equations (2) and (3) (see fig. 1).

The welfare gain of the tax reform ΔU is given by the difference between the utility on the path BE and the utility in the steady state A:[13]

$$\Delta U = J(\bar{k}) - \frac{u(\bar{c})}{\rho^* - n - \mu}. \tag{9}$$

[10] When the tax revenues are not equal to zero, this is equivalent to the gain obtained by a shift from an interest tax to lump-sum taxation.

[11] It is essential throughout this study that the tax changes are unanticipated; for an analysis of the effects of anticipated tax changes, see Hall (1971)

[12] The observed elasticity of gross savings η, with respect to the net rate of return at time zero (under the assumption of perfect foresight about future interest rates), can be determined from the parameters of the model: $\eta = \alpha\epsilon\{r^*/[(r^* + \delta)(n + \mu + \delta)(1 - \alpha)]\}$, where α and δ are defined in n. 7 and a is given in table 2. For the values used below, the Cobb-Douglas case ($\sigma = \epsilon = 1$), $\eta = .95$.

[13] When the tax rate θ is small, the following experiment is symmetrical and gives the same result: Assume that the economy is in the steady state with no tax (point E). At time zero, the tax is instituted (with lump-sum refunds). The economy moves then on a path toward the point A. Call J_1 the utility level on this path and J_0 the utility level at the steady state E. The excess burden of the interest tax is equal to $J_0 - J_1$

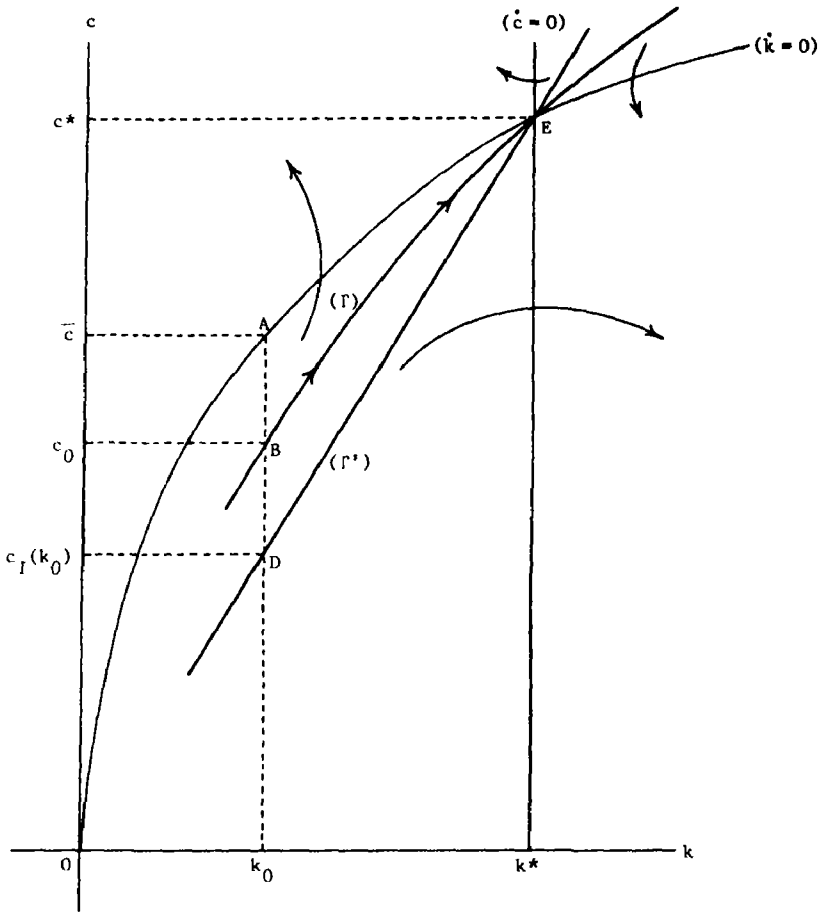


FIG. 1.—The consumption function and the dynamic path

A Taylor expansion of (9) around k^* gives

$$\Delta U \sim J(k^*) + J'(k^*)(\bar{k} - k^*) + \frac{J''(k^*)}{2}(\bar{k} - k^*)^2$$

$$- \frac{u(c^*)}{\rho^* - n - \mu} - \frac{u'(c^*)}{\rho^* - n - \mu}\frac{d\bar{c}}{d\bar{k}}(\bar{k} - k^*)$$

$$- \frac{u''(c^*)}{2(\rho^* - n - \mu)}\frac{d^2\bar{c}}{d\bar{k}^2}(\bar{k} - k^*)^2.$$

From (5a) $d\bar{c}/d\bar{k} = f'(k^*) - n - \mu = \rho^* - n - \mu$. Furthermore, the marginal value of capital $J'(k^*)$ is equal to the marginal utility of

consumption $u'(c^*)$. Therefore,

$$\Delta U \sim \frac{1}{2}\left[f''(k^*) - \frac{u''(c^*)}{\rho^* - n - \mu}\frac{d^2\bar{c}}{d\bar{k}^2}\right]\left(\frac{d\bar{k}}{d\theta}\right)^2\theta^2. \tag{10}$$

This welfare cost is of a second order with respect to the tax rate. It is convenient to divide it by the marginal utility of consumption at the stationary point E, $u'(c^*)$, in order to obtain its wealth equivalent ΔM:[14]

$$\Delta M = L\theta^2\frac{c^*}{r^* - n - \mu}, \tag{11}$$

with

$$L = \frac{1}{2\sigma}\left(\frac{r^*a}{\gamma}\right)^2; \tag{12}$$

the parameters a and γ have been defined in the previous section.

The relation (11) is to be interpreted as follows: The welfare cost of a capital income tax at the (small) rate θ is equivalent to a permanent reduction of consumption on the balanced growth path by a fraction $L\theta^2$. The variable L depends on the technology, the utility function, and the growth rate. We now consider some of its properties.

The relation (12) can be rewritten,

$$L = L_P\left[\frac{(r^* - n - \mu)a}{\gamma}\right]^2, \tag{13}$$

where

$$L_P = \frac{1}{2\sigma}\left(\frac{r^*}{r^* - n - \mu}\right)^2. \tag{14}$$

From (14), and the definition of a and γ (eq. [7]), it is straightforward to derive the following properties:

$$L < L_P;\ \frac{\partial L}{\partial \epsilon} > 0;\ \frac{\partial L}{\partial \sigma} < 0;$$

$$\lim_{\epsilon\to\infty} L = L_P;\ \lim_{\epsilon\to 0} L = 0.$$

The excess burden of the tax is an increasing function of the elasticity of substitution ϵ between capital and labor in the production function. When ϵ is equal to zero, the capital-labor ratio is fixed, and since the labor supply is fixed, there is no distortion. The welfare cost is nil. When ϵ is infinite, the gross factor prices are fixed, $L = L_P$. This

[14] The relations (11) and (12) are derived in the Appendix, Sec. II.

TABLE 1

NORMALIZED EXCESS BURDEN OF THE CAPITAL INCOME TAX

	ϵ										
σ	2	.4	.6	.8	1.0	1.2	1.4	1.6	1.8	2.0	∞
5	1.24	2.41	3 56	4.67	5.76	6.83	7.88	8.91	9 92	10.93	400.0
1.0	1.21	2.33	3.41	4.45	5.46	6.44	7.40	8.33	9.25	10.14	200.0
1 5	1.19	2.28	3.31	4.30	5.25	6.16	7.05	7.92	8.76	9.58	133.33
2.0	1.17	2.23	3.22	4.17	5.07	5.94	6.78	7.59	8.37	9.13	100.0

NOTE —Given by eq (12), in percentage of annual consumption.

TABLE 2

ANNUAL RATE OF CONVERGENCE TOWARD BALANCED GROWTH PATH

	ϵ									
σ	.2	4	6	8	1.0	1.2	1.4	1.6	1.8	2 0
.5	33.98	23.74	19.21	16 51	14 67	13.31	12 25	11.40	10.70	10.10
1 0	23.74	16.51	13.31	11.40	10.10	9.14	8.40	7.80	7.30	6.88
1 5	19 21	13.31	10.70	9.14	8.08	7.30	6.69	6.21	5.80	5.46
2 0	16.51	11.40	9.14	7.80	6.88	6.21	5.68	5.26	4.91	4.62

NOTE —a is measured in percentage

case is analogous to the partial equilibrium situation with exogenous factor prices and provides an upper bound for the general formula.[15]

Since $1/\sigma$ is an index of the intertemporal substitutability of consumption in the utility function, it is not surprising that the excess burden is a decreasing function of σ.

Tables 1 and 2 present estimates of the variable L and of the annual rate of convergence of the economy toward the steady state a for different values of ϵ and σ.[16] The other parameters of the model are chosen to characterize the U.S. economy:[17] $\mu = .02$, $n = .0$, the gross capital income share is equal to .33, the rate of capital depreciation is equal to 5 percent, and ρ^* is equal to the long-run value of the net rate of return and is taken to be equal to 4 percent. There remain only two

[15] This formula has been derived by Levhari and Sheshinski (1972) for a stationary economy. Another proof, using the well-known Harberger-Hicks-Hotelling formula, is given in the Appendix, Sec. I.

[16] These values of the rate of convergence can be compared with those discussed by Sato (1966) and give additional information for a choice of ϵ and σ.

[17] See, e.g., Jorgenson and Christensen 1973.

unknown parameters, ρ and σ. Since there is no general agreement about their values,[18] we present estimates for different values of σ. Once a value is chosen for σ, the pure rate of time preference is implicitly determined by the relation, $\rho = \rho^* - \sigma\mu$.

As can be seen from table 1, for realistic values of ϵ the excess burden is not very sensitive to the elasticity of the marginal utility σ.[19]

The last column in table 1 corresponds to the partial equilibrium case. It is clear that for realistic values of ϵ, the excess burden is much smaller. The variable L is represented as a function of ϵ for a fixed value of σ, equal to one, on figure 2 (curve $\beta = 0$). For the relevant range of ϵ, the excess burden of the interest tax increases almost linearly with ϵ.

The value of the normalized excess burden L depends also on the discount rate r^* and on the growth rate $n + \mu$. No general rule can be derived about the effects of these parameters. When the elasticity ϵ is large, L increases with $n + \mu$ and decreases with r^* (as it is clear from the limit case $\epsilon = \infty$, described by eq. [14]). However, when the elasticity of substitution is smaller than two, and the other parameters of the model have the same values as above, it is found that L is an increasing function of both the discount rate and the growth rate.[20]

The Excess Burden for Large Tax Rates

The measure of the excess burden given by expressions (11) and (12) is exact only for infinitesimal values of the tax rates. In order to estimate the bias involved for large values of θ, we have to use a different method.

At time zero, the economy is on its balanced growth path under taxation (point A on fig. 1), and the initial values of the capital stock and the consumption level are given by (4a) and (5a). The dynamic path after the tax cut, characterized by (2) and (3), is transformed into

[18] Available estimates of σ seem to point to a value higher than one. Weber (1975) reports values between 1.3 and 1.8. Wright's (1969) estimates are somewhat higher—around 4.

[19] In the limit case where the discount rate r^* and the growth rate are equal, the variable L becomes

$$L = -\frac{1}{2}\frac{r^{*2}}{c^* f''(k^*)} = \frac{\alpha}{2(1-\alpha)^2}\cdot\left(\frac{n+\mu}{n+\mu+\delta}\right)^2\cdot\epsilon,$$

which is independent of the utility function and linear in the elasticity ϵ (ceteris paribus).

[20] When $\epsilon = \sigma = 1$, the following values are obtained: Take $n + \mu = 2$ percent; when r^* increases from 2 percent to 6 percent, L increases from 3.00 percent to 6.96 percent. When r^* is equal to 4 percent and $n + \mu$ increases from 2 percent to 4 percent, L increases from 5.46 percent (table 1) to 7.26 percent.

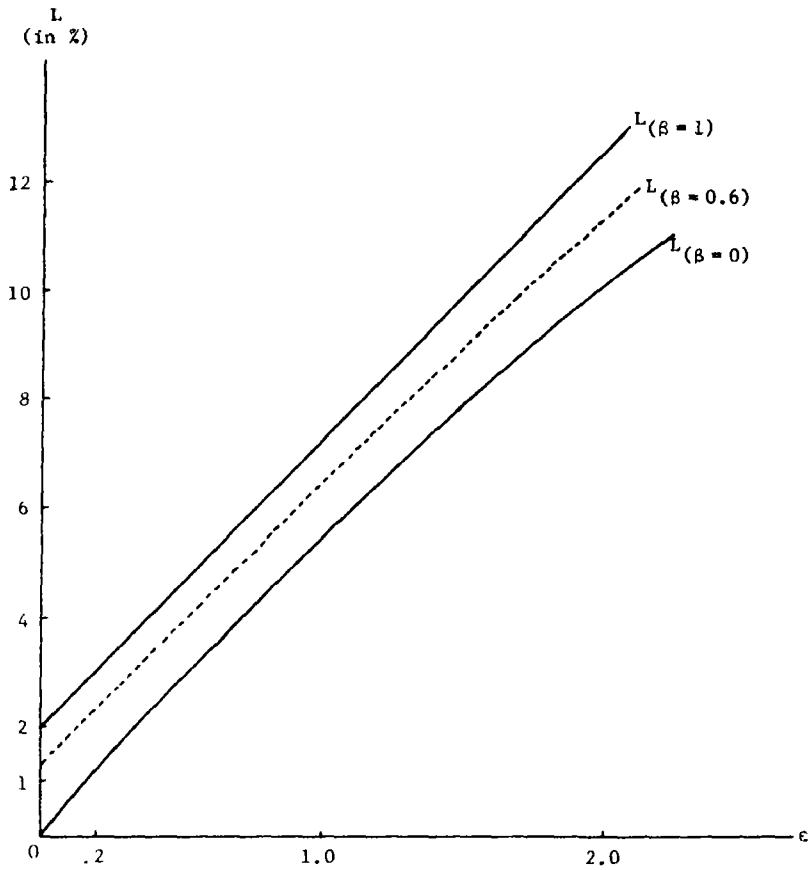


FIG. 2.—The normalized excess burden; ϵ = elasticity of substitution between capital and labor, β = compensated elasticity of substitution between leisure and consumption.

a discrete formulation and simulated numerically using a gradient algorithm of optimal control. The consumption equivalent of the utility after the tax cut c_1 is defined by the stationary level providing the same utility as the dynamic path after the tax cut: $u(c_1)/(\rho^* - n - \mu) = J(\bar{k})$. Table 3 reports the ratio $[(c_1 - \bar{c})/\bar{c}](1/\theta^2)$ as a function of θ for $\sigma = \epsilon = 1$; the other parameters of the model have the same values as in the first two tables.

The first number in the table (for $\theta = 0$) is obtained by formula (12). When θ is small, it gives a good approximation of the normalized excess burden. However, we can see that for large values of the tax rate θ, the approximation formula (12) underestimates the welfare

TABLE 3

NORMALIZED EXCESS BURDEN FOR LARGE TAX RATES (%)

	θ						
	.0	.05	.1	.2	.3	.4	.5
$\frac{c_1 - \tilde{c}}{\tilde{c}} \cdot \frac{1}{\theta^2}$	5.46	5.82	6.25	7.27	8.58	10.33	12.77

NOTE.—$\sigma = \epsilon = 1$

gain of the tax cut (for $\theta = 50$ percent, by a factor slightly smaller than $\frac{1}{2}$).[21]

In order to illustrate the previous discussion, let us consider two numerical examples of a shift from capital income taxation to lump-sum taxation.

Assume that the aggregate consumption in the steady state under taxation is normalized to one. When the tax rate θ is equal to 50 percent, total revenues are equal to 12.34 percent. A shift to a lump-sum tax increases the level of welfare by a consumption equivalent of 3.19 percent and the level of consumption in the long run by 8.38 percent. The welfare cost of the capital income tax is equal to 26 percent of the amount of tax revenues. The marginal welfare cost is obtained by multiplying the average cost by a factor of two (this factor should be somewhat greater than two for large rates, since the normalized excess burden increases with the tax rate).

When θ is equal to 30 percent, tax revenues are equal to 6.73 percent. A shift to lump-sum taxation would increase consumption in the long run by 3.3 percent. The welfare cost of the tax is equal to 0.77 percent of the consumption level or 11 percent of the revenues.

The method followed in this section takes into account both the short-run and the long-run incidence of the suppression of the capital income tax. We can observe that the welfare gain of tax reform (measured in consumption equivalent) is much smaller than the percentage increase of consumption in the long run. Also, the excess burden of the capital income tax, although small in terms of aggregate consumption, is not negligible in terms of the revenues generated.

III. The Case of the Nonperfect Foresight

In the previous section, the welfare gain of a reform from capital income taxation to lump-sum taxation was determined under the

[21] On the other side, when θ is large, the model considered here implies a sudden increase of saving which could be too large to be borne by a real economy. Therefore, it is likely that the numbers in table 3 may be overestimated.

assumption that the economy follows an optimal path toward the new steady state. However, when there is no complete set of future markets to convey information on future prices, individuals may not have perfect foresight. In this case, the dynamic path after the tax reform is no longer optimal, and the welfare gain obtained is smaller.[22]

The assumption of perfect foresight is relaxed in this section. For its saving decision, the competitive household relies now on point expectations about the future wage and interest rates. By assumption, these expectations satisfy the following property: At each instant t, anticipated future prices depend only on the value of the capital stock at time t (e.g., through the factor prices at time t) and on some constant parameters (as the long-run values of the factor prices).[23] It follows that the consumption level can be expressed as a function of k, $c_l(k)$.

We also assume that expectations are consistent with the steady state and do not affect its stability. The rate of convergence of the economy a_l depends on the type of expectations considered. For example, in the case of myopic expectations (where the factor prices observed at a given time are expected to be the same in the future), the rate of convergence is greater than in the case of perfect foresight.

As in the previous section the initial situation of the economy is the steady state with taxation. After the suppression of the tax at time zero, the economy converges to the steady state E on the path $\Gamma'(B'E)$, which in general is different from the path Γ. The welfare gain is measured by the difference,

$$\Delta U = \bar{J}(\bar{k}) - \frac{u(\bar{c})}{r^* - n - \mu}, \tag{15}$$

where $\bar{J}(\bar{k})$ represents the integral (1) on the path $B'E$.

The second-order approximation of the utility level after the tax cut, $\bar{J}(\bar{k})$, depends only on the slope of the dynamic path at the steady state,[24] $c_l'(k^*)$, or on the rate of adjustment toward the steady state a_l (since $a_l = c_l'[k^*] - [r^* - n - \mu]$):

$$\bar{J}(\bar{k}) = F(a_l, \bar{k}); \tag{16}$$

this utility is smaller than the utility on the optimal path,

$$F(a_l, \bar{k}) \leq F(a, \bar{k}) = J(\bar{k}). \tag{17}$$

[22] For the class of expectations considered below, the dynamic path still converges to the steady state.

[23] This class includes perfect foresight and myopic, stationary, and regressive expectations. This section uses some intuitive results derived in Chamley (1979).

[24] A proof is given in the Appendix, Sec. III.

The welfare gain induced by the tax cut is equivalent to a permanent increase of the consumption level by a fraction $L_l\theta^2$, where L_l is defined by

$$L_l = [Q(a_l)/Q(a)] \cdot L, \tag{18}$$

with

$$Q(a_l) = \frac{a_l}{\lambda + 2a_l}\left[\frac{a_l}{a \cdot c'(k^*)} - \frac{2}{\lambda}\right]; \tag{19}$$

L is equal to the perfect-foresight value defined by (12) in the previous section.

Using the relations (18) and (19) and the properties of a and $c'(k^*)$ (described in Sec. I), it is straightforward to show that the ratio L_l/L is always smaller than one. When the private sector does not anticipate future prices with perfect foresight, the welfare gain induced by the suppression of the capital income tax is reduced.

The ratio between the values of the excess burden with nonoptimal and optimal adjustments L_l/L is represented in figure 3 as a function of the ratio between the rates of convergence (a_l/a). The different curves correspond to different values of the discount rate r^*. The growth rate $n + \mu$ is equal to 2 percent, and the product $\sigma\epsilon$ is equal to 1.

Also the cases $r^* = 3$ percent and $r^* = 6$ percent provide a very good approximation (less than 1 percent error) of the cases $r^* = 4$ percent, $\sigma\epsilon = \frac{1}{2}$, and $r^* = 4$ percent, $\sigma\epsilon = 2.5$, respectively.

It is interesting to observe that, in general, a shift from capital income taxation to lump-sum taxation always induces a welfare gain (L_l is positive unless a_l has an unrealistically large value, greater than $[12.1] \cdot a$). This gain is close to its maximum for a wide range of values of the ratio (a_l/a). When this ratio is between $\frac{1}{3}$ and 3, L_l is equal to at least 86 percent of L (for $r^* = 4$ percent, $\sigma\epsilon = 1$).

As an example, let us consider the case of myopic expectations. The consumption function takes then the following form:

$$c_S(k) = \left(1 + \frac{1}{\sigma}\,\frac{\rho^* - r}{r - n - \mu}\right)[f(k) - (n + \mu)k].$$

Using (18) and (19), after elementary manipulations, the ratio between the normalized excess burden under myopic expectations and its value under perfect foresight, L_S/L, is given by

$$L_S/L = \frac{1}{2}\left(1 + \frac{\lambda\sqrt{\lambda^2 + 4\gamma}}{\lambda^2 + 2\gamma}\right), \tag{20}$$

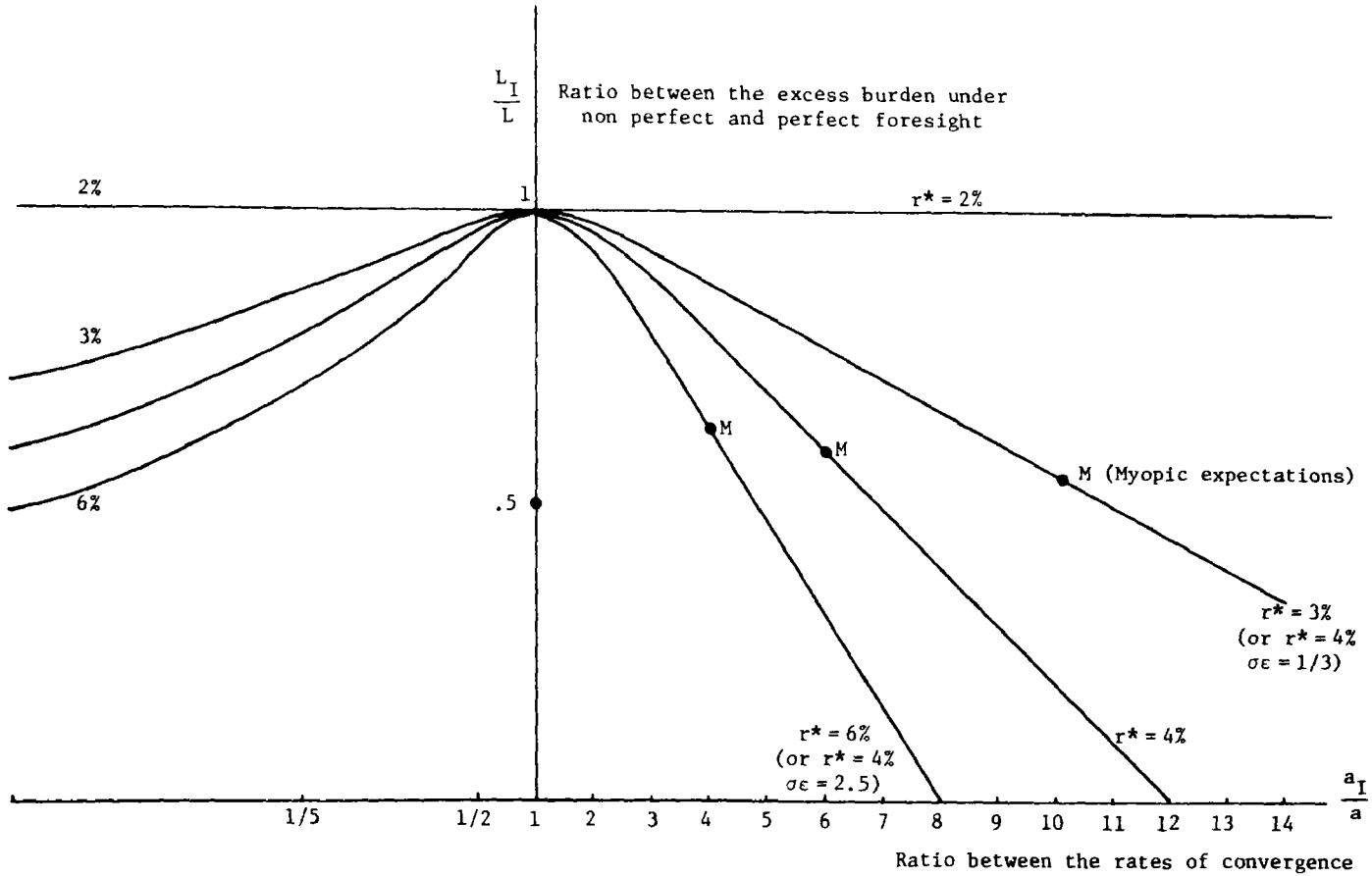


FIG. 3.—The excess burden under nonperfect foresight

with $\lambda = r^* - n - \mu$, and γ is the parameter described in equation (7). A good approximation of (20) is given by

$$L_S/L = \frac{1}{2}\left(1 + \frac{\lambda}{\lambda + a}\right), \tag{21}$$

where a is the optimal rate of convergence reported in table 2.

Under myopic expectations, the welfare gain of the tax reform is always equal to at least 50 percent of the perfect-foresight value. This case is represented by the point M on figure 3 (for $\epsilon = \sigma = 1$, the rate of convergence toward the steady state under myopic expectations a_S is about six times larger than the value under perfect foresight a). The ratio L_S/L depends on the long-run value of the rate of return r^*, the growth rate $n + \mu$, and the parameters ϵ and σ. When $r^* = 0.04$ and $n + \mu = 0.02$, it can be determined for different values of ϵ and σ, using table 2 and the relation (21).

We see in figure 3 that, for the relevant range of values of the various parameters, the ratio between the welfare gains of tax reform under myopic expectations and under perfect foresight is between .55 and .65 and is not very sensitive to the parameters of the model.[25]

Figure 3 also allows us to consider a more general class of expectations. For example, when the private sector underestimates the future rate of convergence of the economy, the realized rate of convergence a_I is in the interval (a, a_S). For this special class of expectations, the welfare gain of a tax cut is bounded below by the value found in the myopic case.

Finally, it may be interesting to observe that when the private sector applies a modest amount of rationality, the dynamic path of the economy is close to the optimal path. Around the steady state, the variation through time of the factor prices is given by the expression, $(\dot{z}_t - z^*) = -a_I(z_t - z^*)$.

Assume now that the long-run values of the factor prices are known and that the only uncertainty left is about their rate of convergence to the new steady state. Furthermore, individual expectations are of the regressive form; the expected values of a parameter x, at some future date t, x_t^e, are determined by the following rule:

$$(\dot{x}_t^e - x^*) = -\alpha(x_t^e - x^*), \tag{22}$$
$$x_0^e = x_0.$$

Individuals revise at each instant the estimated value α of the future rate of convergence, using the observed value $\hat{a}$ (given by $\hat{a} = \dot{z}/(z^* -$

[25] When the difference λ between the rate of return r^* and the growth rate $n + \mu$ tends to zero, the ratio L_S/L tends to 1/2. Also, at the same time, the error implied by myopic expectations, which is measured by the ratio a_S/a, tends to infinity.

z), where z is an arbitrarily chosen endogenous variable) and an adaptive rule, for example, $\dot{\alpha} = \nu(\hat{a} - \alpha)$, where $\nu > 0$.

Under these assumptions, and for an initial value of the capital stock not too different from the steady-state value, both the estimated value α and the actual value $\hat{a}$ of the rate of convergence tend to the optimal value a.[26] The dynamic path is *tangent* to the perfect foresight path at the steady-state point. The value of the welfare gain of tax reform is close to its (maximum) perfect-foresight value.

IV. The Welfare Cost of the Corporation Tax

Most studies on the corporation tax assume that the corporate and the noncorporate sectors produce two different goods and that the ratio between their respective outputs depends on their relative prices.[27] For the sake of simplicity, we assume here that both sectors produce the same good and that the aggregate production function can be written in the form,[28] $y = g(k_1, k_2)$, where all quantities are divided by the total effective labor supply, and k_1 and k_2 represent the corporate and the noncorporate capital stock, respectively. The function $g(k_1, k_2)$ is assumed to be homogeneous in its arguments.

Initially the economy is in the steady state where only the corporate capital income is taxed at the rate θ. In each sector, the net rate of return is equal to the optimal stationary value ρ^*, and the values of the capital stocks, $\bar{k}_1$ and $\bar{k}_2$, are determined by

$$\rho^* = (1 - \theta)\frac{\partial g}{\partial k_1}(\bar{k}_1, \bar{k}_2) = \frac{\partial g}{\partial k_2}(\bar{k}_1, \bar{k}_2). \tag{23}$$

The elimination of the corporate tax at time zero has two effects. First, capital is reallocated between the two sectors to equalize their

[26] Numerical simulations have shown that for almost any initial value of the capital stock, the convergence to the perfect-foresight path occurs after only a few periods. This procedure may be used as an algorithm to determine the perfect-foresight solution.

[27] See Harberger (1962, 1976), Shoven and Whalley (1972), Boadway and Treddenick (1975), and Shoven (1976) for a static analysis. Friedlander and Vandendorpe (1978) extend Harberger's study to a dynamic framework. However, they do not address the problem of the incidence on the intertemporal welfare and do not consider a saving function derived from the optimization of an intertemporal utility function. Fullerton, Shoven, and Whalley (1979) are currently working on a more elaborate model.

[28] The usual assumption that the corporate and the noncorporate sectors produce two separate goods may also be a crude description of reality; see, e.g., Ebrill and Hartman 1977.

rates of return. The aggregate production function now takes the form,

$$y = f(k), \tag{24}$$

where

$$f(k) = \max_{k_1+k_2=k} g(k_1, k_2).$$

After time zero, the private sector is endowed with perfect foresight, and the economy moves on the dynamic path studied in Section I, which is characterized by the equations, $\dot{c} = (c/\sigma)[f'(k) - \rho^*]$, $\dot{k} = f(k) - (n + \mu)k - c$, and $k_0 = \bar{k} = \bar{k}_1 + \bar{k}_2$. As in the relation (9), the welfare gain of the tax cut is equal to $\Delta U = J(\bar{k}) - [u(\bar{c})/(r^* - n - \mu)]$. This difference can be decomposed into

$$\Delta U = \left[J(\bar{k}) - \frac{u(\hat{c})}{r^* - n - \mu}\right] + \left[\frac{u(\hat{c}) - u(\bar{c})}{r^* - n - \mu}\right], \tag{25}$$

where $\hat{c} = f(\bar{k}) - (n + \mu)\bar{k}$ is the stationary consumption available when the level of capital $\bar{k}$ is allocated efficiently between the corporate and the noncorporate sectors.

The first term of the RHS (right-hand side) of (25) has been analyzed in Section II and measures the intertemporal welfare cost of the corporate tax, originating in the reduction of capital accumulation. It is equivalent to a permanent reduction of the consumption level by a fraction $L_{C1}\theta^2$, where L_{C1} is defined by[29]

$$L_{C1} = \left(\frac{\bar{k}_1}{\bar{k}}\right)^2 \cdot L. \tag{26}$$

Since the tax revenues are equal to $\theta r^* k_1$, this relation implies that for small values of the tax rates, the intertemporal welfare cost of capital income taxation depends only on the total amount of tax revenues (e.g., a 10 percent uniform tax on the whole capital stock and a 20 percent tax on half the capital stock have up to the second order the same intertemporal welfare cost).[30]

Although the distribution of the tax burden across the different sectors does not affect the intertemporal welfare cost, it creates a distortion in the allocation of the capital stock between sectors. This

[29] A proof is given in the Appendix, Sec. IV.

[30] The extension to many sectors is straightforward. In particular, this analysis could be applied to a model with four types of capital: corporate and noncorporate capital, housing, and human capital.

production cost is measured by the second term in the RHS of (25) and is equal to[31]

$$\frac{u(\hat{c}) - u(\bar{c})}{(r^* - n - \mu)} = \frac{u'(\bar{c})}{(r^* - n - \mu)} [f(\bar{k}) - g(\bar{k}_1, \bar{k}_2)]. \tag{27}$$

The welfare cost induced by the production inefficiency is equivalent to a permanent reduction of the consumption level by a fraction $L_{C2}\theta^2$, where

$$L_{C2}\theta^2 = [f(\bar{k}) - g(\bar{k}_1, \bar{k}_2)]/\bar{c}. \tag{28}$$

This is precisely the excess burden studied by Harberger and others.

The total welfare cost of the corporate tax is equal to the sum of the intertemporal cost and of the cost due to production inefficiency. It is equivalent to a permanent reduction of the consumption level by a fraction $L_C\theta^2$, where

$$L_C = L_{C1} + L_{C2} = L\left(\frac{\bar{k}_1}{\bar{k}}\right)^2 + \frac{[f(\bar{k}) - g(\bar{k}_1, \bar{k}_2)]}{\bar{c}}; \tag{29}$$

$\bar{k}_1$, $\bar{k}_2$, and $\bar{k}$ are given by (21), and

$$\bar{c} = f(\bar{k}) - (n + \mu)\bar{k}. \tag{30}$$

In order to have an order of magnitude of the respective quantities, assume that the ratio k_1/k is equal to $\frac{1}{3}$, which is close to the value observed in the U.S. economy, and that only the capital income originating in the corporate sector is taxed at the rate of 50 percent. Using the results of Section II, the intertemporal welfare cost is about 0.25 percent of the level of aggregate consumption (or around 7.8 percent of the tax revenues).[32] The value is somewhat lower than the cost of intersectoral misallocation, which in the same model is equal to 1.1 percent of consumption (this value is well within the range of current estimates of 0.5–1.5 percent; see Shoven 1976). The interindustry misallocation caused by the corporate tax seems to be much greater than the intertemporal distortion.

V. The Case of an Elastic Labor Supply

The capital income tax increases the gross rate of return and, by the factor price frontier, lowers the wage rate. When the labor supply is

[31] To a second order, it is equivalent to evaluate this difference at $\bar{k}$ or k^*.

[32] These estimates are obtained with a simplified version of the model in Section II: There is no depreciation, and the production function takes the form $y = k_1^{\alpha}k_2^{\beta}$ (with a 50 percent corporate tax rate, $k_1/k_2 = 0.5$). The capital income share, net of depreciation, 2β, is the same as for the previous numerical results and is equal to 0.18. We have also used table 3 in the correction for large tax rates.

not fixed, this tax creates, in addition to the intertemporal distortion, a distortion in the choice between consumption and leisure at a given instant of time.

For example, assume that the elasticity of substitution between capital and labor in the production function is equal to zero; the capital-labor ratio is constant. When the labor supply is fixed, the tax has no incidence on the capital stock and is equivalent to a lump-sum tax with no excess burden. However, in the long run the gross rate of return is higher and, by the price possibility frontier, the wage rate is lower. When the labor supply depends on the net factor prices, the incidence of the tax is to decrease the long-run labor supply, the capital stock, and the consumption level.[33]

In general, it can be expected that the excess burden of the capital income tax is increased when the labor supply is variable. In this section, the framework developed in Section II is extended in order to measure this additional effect.

The utility function of the private sector depends on the amounts of consumption and leisure:

$$U = \int_0^\infty e^{-\rho t} e^{nt} u(c_t e^{\mu t}, l_t) dt, \tag{31}$$

where $c_t e^{\mu t}$ is the consumption per capita and l_t is the labor supply per capita (measured in natural units, and not in effective units).

For the sake of simplicity, we consider the following form for the utility function u:[34] $u(c,l) = (1 - \beta) \log c + \beta \log (T - l)$, where β and T are exogenous parameters. The dynamic path of the economy is now characterized by the relations,

$$\dot{c}_t = c_t[r_t(1 - \theta) - \rho^*], \tag{32}$$

$$\dot{k}_t = f(k_t, l_t) - (n + \mu)k_t - c_t, \tag{33}$$

and

$$\frac{\beta c_t}{(1 - \beta)(T - l_t)} = w_t, \tag{34}$$

where r_t and w_t are the gross interest and wage rates, respectively, and k_t represents now the aggregate capital stock divided by the efficiency index $e^{\mu t}$.

Equations (32) and (33) are the same as in the fixed labor supply case, and, at each time t, the marginal rate of substitution between the

[33] A sufficient condition for these properties to be verified is the homotheticity of the utility function. This homotheticity is a necessary condition for the existence of an optimal balanced growth path.

[34] When μ is different from 0, and the elasticity of substitution in u between consumption and leisure is different from 1, there is no optimal balanced growth path.

consumption of leisure and of produced goods is equal to the wage rate (eq. [34]).

At a given instant, the labor supply depends not only on the wage rate at the same moment but also on the future factor prices (wage and interest rates).

The initial state of the economy (where the tax has been in effect for an infinite amount of time) is defined by the stationary equivalents of (32)–(34):

$$\bar{r} = \frac{\rho^*}{1 - \theta}, \tag{35}$$

$$\bar{c} = A(\bar{r})\bar{l}, \tag{36}$$

and

$$\frac{\beta\bar{c}}{(1 - \beta)(T - \bar{l})} = w(\bar{r}), \tag{37}$$

where $A(r)$ is defined using the production function, $A(r) = r[(k/l)(r)] + w(r) - n - \mu$, and the capital-labor ratio k/l is expressed as a function of the gross rate of return r.

At time zero, the tax is suppressed. Thereafter, the dynamic behavior of the economy is described by (32), (33), and (34), where θ is replaced by 0; the new steady state is characterized by

$$r^* = \rho^*, \tag{38}$$

$$c^* = A(r^*)l^*, \tag{39}$$

and

$$\frac{\beta c^*}{(1 - \beta)(T - l^*)} = w(r^*). \tag{40}$$

The increase of welfare induced by the tax cut can be approximated to the second order by the same method as in Section II; its wealth equivalent is of the form, $\Delta M = L_\beta\,[c^*/(r^* - n - \mu)]\theta^2$, which depends on the parameter β. The formula for L_β is rather complicated,[35] so we present in tables 4 and 5 some numerical estimates analogous to those in Section II.

Also, L_β is represented as a function of ϵ in figure 2, for the values $\beta = 0, .6, 1$; β measures the long-run compensated elasticity of the labor supply with respect to the wage rate. When it is equal to zero, the labor supply is fixed; this is the case studied in Section II.

In table 4 and figure 2, we can see that the excess burden $L_\beta\theta^2$, for a given value of θ, is increasing with β and is bounded by the value of

[35] A computer program for its numerical evaluation is available.

TABLE 4

NORMALIZED WELFARE COST OF CAPITAL INCOME TAX WITH A VARIABLE LABOR SUPPLY (%)

	ϵ										
β	.0	.2	.4	.6	8	1 0	1.2	1.4	1.6	1.8	2.0
.0	.0	1.21	2 33	3.41	4.45	5.46	6.44	7.40	8.33	9.25	10.14
.2	.45	1.60	2.71	3.77	4.81	5.82	6.80	7.76	8.71	9.63	10.54
.4	.86	1.98	3.07	4.13	5 16	6.17	7.16	8.13	9.09	10.03	10.96
.6	1.25	2.34	3.42	4 47	5.51	6.53	7.53	8.52	9.50	10.47	11.42
.8	1.61	2.68	3.75	4.81	5.85	6.89	7.92	8.93	9.94	10.95	11.94
1.0	1.95	3.01	4.07	5.13	6.20	7.26	8.32	9.39	10.45	11.51	12.58

TABLE 5

ANNUAL RATE OF CONVERGENCE TOWARD BALANCED GROWTH PATH WITH A VARIABLE LABOR SUPPLY (%)

	ϵ										
β	.0	.2	.4	.6	.8	1.0	1.2	1.4	1.6	1.8	2.0
.0	∞	23.74	16 51	13.31	11.40	10 10	9.14	8.40	7.80	7.30	6.88
.2	40.09	22.17	16.85	14.06	12.27	11.00	10.04	9.28	8.66	8.14	7.70
.4	28.47	20.91	17.20	14.90	13.31	12.11	11.18	10.42	9.79	9 25	8.79
.6	23.35	19.87	17.55	15.86	14.57	13.53	12.68	11.96	11.35	10.82	10.35
8	20.32	19.01	17.91	16.97	16.16	15.45	14.83	14.27	13.76	13.30	12.89
1.0	18.28	18.28	18.27	18.27	18.26	18.26	18.25	18.24	18.24	18.23	18.23

$L_1\theta^2$, obtained for $\beta = 1$.[36] It is remarkable that the difference $L_\beta - L_0$ does not appear to be significantly affected by the value of ϵ[37] (for ϵ smaller than 2) and that this difference is relatively small with respect to L_0 when the production function is of the Cobb-Douglas type ($\epsilon = 1$).

When ϵ is infinite, we have the partial equilibrium case again. It is straightforward to show that the relation (14) is still valid.

The welfare cost of the capital income tax at the rate θ is measured by a fraction of the full consumption of produced goods and of leisure (valued at the wage rate), which is equal to $L_P\theta^2$, where $L_p = \frac{1}{2}[r^*/(r^* - n - \mu)]^2$.

VI. Conclusion

The general equilibrium models used in this study are highly stylized. However, the numerical examples considered indicate that the character of its results is fairly general.

The value of the excess burden of the capital income tax depends mainly on the elasticity of substitution between capital and labor ϵ. When ϵ is smaller than two, the excess burden is almost proportional to ϵ. No general rule could be derived about the effects of the discount rate. When the other parameters of the model are chosen to characterize the U.S. economy, the excess burden of the capital income tax increases with the discount rate and the growth rate.

Interestingly enough, when ϵ is smaller than two, an increase of the intertemporal substitutability of consumption (measured here by the inverse of the marginal utility of consumption) increases the value of the excess burden only by a negligible amount. In this case, although the deadweight loss of the tax is relatively small with respect to the level of aggregate consumption, or to its value in the case of fixed factor prices ($\epsilon = \infty$), it is not negligible with respect to the amount of tax revenues: A capital income tax of 50 percent implies a welfare loss equivalent to 26 percent of the tax revenues; when the tax rate is equal to 30 percent, the deadweight loss is equal to 11 percent of the tax revenues. These values should be multiplied by a number between 1 and 4/3 if the labor supply is elastic.

These numbers are indicative of the welfare gain induced by a shift from capital income taxation to lump-sum taxation, when individuals

[36] The compensated elasticity of the labor supply β is bounded by 1.

[37] Numerical experiments have shown that the difference $L_1 - L_0$ is not sensitive to variations of δ, ρ, or μ, for realistic values of ϵ. When ϵ is equal to zero, the excess burden of the capital income tax arises because of its incidence on the wage rate. In this special case, the interest tax has the same excess burden as the wage tax which generates an equal amount of revenues (Chamley 1980*a*).

have perfect foresight about the behavior of the economy after the tax is suppressed. Since the determination of the perfect-foresight path is not an easy task for most economists, even in a simple model, this assumption may not be realistic. When the tax reform does not lead to an optimal path, its benefits are reduced. For example, under myopic expectations, the welfare gain induced by the abolition of the capital income tax is only equal to about 60 percent of its value under perfect foresight. However, myopic expectations grossly underestimate the future behavior of the economy. When individuals anticipate somewhat this future behavior, the welfare gain of tax reform is close to its maximum value, even if the degree of foresight is not very accurate.[38] (A method to approximate the perfect foresight path has also been suggested in Sec. III.)

Finally, under the present U.S. tax system, the rates of the capital income tax vary by a large amount from one sector of production to another. When the overall elasticity of substitution between capital and labor is not too large (less than two), the welfare gains obtained by an equalization of these rates dwarf the gains obtained by a reduction of the global tax on capital income.[39]

Appendix

I. *The Excess Burden in Partial Equilibrium*

An income equivalent of the excess burden is given by the well-known Harberger-Hicks-Hotelling formula. In a continuous-time formulation, this becomes

$$\Delta M = \frac{1}{2}\int_0^\infty \int_0^\infty \Delta p_t \Delta p_{t'} \left(\frac{\partial c_t}{\partial p_{t'}}\right)^U e^{mt} dt dt', \tag{A1}$$

with the following notation:

$$\begin{cases} m = n + \mu \\ p_t = e^{-r^* t(1-\theta\mu)} \text{ (price of } c_t \text{ at time 0)}, \\ \Delta p_t = r^* \theta t e^{-r^* t} \end{cases}$$

$$\left(\frac{\partial c_t}{\partial p_{t'}}\right)^U, \text{ compensated derivative of } c_t \text{ with respect to } p_{t'}.$$

[38] The assumption of myopic expectations is used in Hudson and Jorgenson (1976) and in Fullerton et al. (1979). Both studies rely on multisectoral models where one type of capital is allocated between the different sectors of production. Because there is only one state variable (the aggregate capital stock), regressive expectations are still asymptotically correct, and the discussion in Sec. III applies; in particular, the mechanism suggested there could be used to improve the accuracy of foresight.

[39] This result is also found in Fullerton et al. (1979).

By application of Slutsky's equation,

$$\Delta M = \frac{1}{2}\int_0^\infty e^{mt}\Delta p_t \left(\int_0^\infty \frac{\partial c_t}{\partial p_{t'}} \Delta p_{t'} dt'\right) dt$$

$$+ \frac{1}{2}\int_0^\infty e^{mt}\Delta p_t \frac{c_t}{M}\left(\frac{M}{c_t}\frac{\partial c_t}{\partial M}\right)\left(\int_0^\infty e^{mt'} c_{t'}\Delta p_{t'} dt'\right) dt.$$

The first expression in parentheses is simply equal to the variation of the consumption at time t, Δc_t, after an *uncompensated* interest tax has been applied (at time zero).

Since the utility function is homogeneous, the income elasticity of consumption is equal to one, and the excess burden is given by

$$\Delta M = \frac{1}{2}\int_0^\infty e^{mt}\Delta p_t \Delta c_t dt + \frac{1}{2M}\left(\int_0^\infty e^{mt} c_t \Delta p_t dt\right)^2.$$

From (2) and the budget constraint, the uncompensated demand function, c_t, is expressed as follows: $c_t = \nu M e^{\alpha t}$, with $\nu = r^* - n - \mu - \theta r^*[1 - (1/\sigma)]$, and $\alpha = -(r^*/\sigma)\theta$. Therefore, when θ is small,

$$\Delta c_t = -Mr^*\left[\left(1 - \frac{1}{\sigma}\right) + \frac{(r^* - n - \mu)t}{\sigma}\right].$$

We replace Δc_t in the above expression for ΔM, and since the wealth M is equal to the present value of the consumption stream on the balanced growth path, we have

$$\Delta M = -L_P \frac{c^*}{r^* - n - \mu}\theta^2, \tag{A2}$$

with

$$L_P = \frac{1}{2\sigma}\left(\frac{r^*}{r^* - n - \mu}\right)^2.$$

For a given value of the interest tax rate, the welfare cost in partial equilibrium is an increasing function of the intertemporal substitutability of consumption in the utility function, which is measured by $1/\sigma$. It is a decreasing function of the discount rate and an increasing function of the growth rate.

II. *The Excess Burden in General Equilibrium*

We determine here a second-order equivalent of the welfare gain ΔU, obtained by the suppression of the interest tax: $\Delta U = J(\bar{k}) - [u(\bar{c})/\lambda]$ ($\lambda = \rho^* - n - \mu$). The terms $J(\bar{k})$ and $u(\bar{c})$ are considered separately.

The term $J(\bar{k})$ is equal to

$$J(\bar{k}) = \int_0^\infty e^{-\lambda t} u(c_t) dt,$$

where c_t is taken on the dynamic path BE (fig. 1). This integral can be decomposed into two terms:

$$J = u(c_t)\Delta t + \int_{\Delta t}^\infty e^{-\lambda t} u(c_t) dt,$$

where Δt is a small interval of time. By taking an infinitesimal value for Δt, we verify that the function J satisfies the relation, $0 = u(c) - \lambda J + J'(k)k$; or, using the capital accumulation equation (eq. [3]), $0 = u[c(k)] - \lambda J(k) + J'(k)\{f(k) - (n + \mu)k - c(k)\}$.

Differentiating this expression twice at the point k^*, we obtain

$$J'(k^*) = u'(c^*),\ J''(k^*) = \frac{u''c'^2 + J''u'}{\lambda + 2a},$$

where $u' = (du/dc)(c^*)$, $u'' = (d^2u/dc^2)(c^*)$, $c' = (dc/dk)(k^*)$, $c'' = [d^2c(k^*)]/dk^2$, and $a = c'(k^*) - \lambda$.

Therefore, a second-order approximation of J at $\bar{k}$ is given by

$$J(\bar{k}) = \frac{u}{\lambda} + u'(\bar{k} - k^*) + \frac{1}{2(\lambda + 2a)}(u''c'^2 + J''u')(\bar{k} - k^*)^2. \tag{A3}$$

Using the properties of the optimal consumption function described in Section I of the text, we can rearrange the term of the second order:

$$\frac{u''c'^2 + J''u'}{2(\lambda + 2a)} = \frac{u''}{2(\lambda + 2a)}\left(c'^2 - \frac{cf''}{\sigma}\right)$$

$$= \frac{u''(c'^2 + c'a)}{2(\lambda + 2a)} = \frac{u''c'(c' + \sigma a)}{2(\lambda + 2a)} = \frac{u''c'}{2}.$$

Hence,

$$J(\bar{k}) = \frac{u}{\lambda} + u'(\bar{k} - k^*) + \frac{u''c'}{2}(\bar{k} - k^*)^2. \tag{A4}$$

The value of consumption in the steady state under taxation, $\bar{c}$, is determined by $\bar{c} = f(\bar{k}) - (n + \mu)\bar{k}$. The second-order equivalent of $u(\bar{c})/\lambda$ can be written as

$$\frac{u(\bar{c})}{\lambda} = \frac{u}{\lambda} + u'(\bar{k} - k^*) + \frac{1}{2\lambda}(u'f'' + u''\lambda^2)(\bar{k} - k^*)^2, \tag{A5}$$

$$= \frac{u}{\lambda} + u'(\bar{k} - k^*) + \frac{u''}{2}\left(\lambda + \frac{c'a}{\lambda}\right)(\bar{k} - k^*)^2. \tag{A6}$$

Taking the difference between $J(\bar{k})$ and $u(\bar{c})/\lambda$ (in [A4] and [A6]), we have

$$\Delta U = -\frac{u''}{2}\frac{a^2}{\lambda}(\bar{k} - k^*)^2,$$

where a is the positive root of $x^2 + \lambda x - \gamma = 0$ (described in text Sec. I).

Since $f'(\bar{k}) = \rho^*/(1 - \theta)$ and $(\bar{k} - k^*)f'' \sim \rho^*\theta$, the welfare change can be rewritten as

$$\Delta U = \frac{u'}{2\sigma}\frac{c}{\lambda}\left(\frac{\rho^*}{\lambda}\right)^2\left(\frac{\sigma\lambda a}{cf''}\right)^2\theta^2 = u'\frac{c}{r^* - n - \mu}\theta^2 L,$$

where

$$L = L_P\left(\frac{\lambda a}{\gamma}\right)^2,$$

$$L_P = \frac{1}{2\sigma}\left(\frac{\rho^*}{\lambda}\right)^2$$

(the partial equilibrium value of L expressed by the relation [14] in the text), and

$$\gamma = \frac{-cf''(k^*)}{\sigma}.$$

The term $\lambda a/\gamma$ is equal to the positive root of the equation $(\gamma/\lambda^2)x^2 + x - 1 = 0$. This root is contained in the interval (0,1). The derivation of its properties when ϵ varies is straightforward. The sign of $\partial L/\partial\sigma$ is obtained by the same method.

III. *The Case of Nonperfect Foresight*

When the private sector is not endowed with perfect foresight, $c'(k^*)$ is no longer the positive root of the equation $x^2 - \lambda x - \gamma = 0$. The welfare gain of the interest tax cut is obtained by the same method as before. However, taking the difference between (A3) and (A5), we find

$$\Delta U = u'\left(\frac{\bar{k} - k^*}{2}\right)^2\left[f''\left(\frac{1}{\lambda + 2a_l} - \frac{1}{\lambda}\right) + \frac{\sigma}{c}\left(\lambda - \frac{c_l'^2}{\lambda + 2a_l}\right)\right],$$

where c_l and a_l represent the consumption function and the coefficient of adjustment of the economy toward the balanced growth path, under nonperfect foresight.

Use the relation $a_l = c_l' - \lambda$ and the definition of the perfect-foresight values for a_l and c_l', respectively a and c', to obtain

$$\Delta U = u'(c^*)f''(k^*)\left(\frac{\bar{k} - k^*}{2}\right)^2 Q(a_l),$$

where

$$Q(a_l) = \frac{a_l}{\lambda + 2a_l}\left(\frac{a_l}{ac'} - \frac{2}{\lambda}\right).$$

The consumption equivalent of the welfare gain can then be expressed as $L_l = L \cdot Q(a_l)/Q(a)$, where L is the value of L_l under perfect foresight.

IV. *The Corporation Tax*

We examine here the long-run incidence of the corporation tax on the capital stock. The capital stocks in the corporate and in the noncorporate sector, k_1 and k_2, respectively, are determined in the long run by the relations $\rho^* = (1 - \theta)g_1'(k_1, k_2) = g_2'(k_1, k_2)$. Differentiate this expression around $\theta = 0$ to obtain the variation of the aggregate capital stock, dk:

$$dk = dk_1 + dk_2 = \left(\frac{g_{22}'' - g_{12}''}{\Delta}\right)\rho^*\theta, \tag{A7}$$

with $\Delta = g_{11}''g_{22}'' - g_{12}''g_{21}''$. Assume now that the tax is applied uniformly to both sectors, and call Dk the variation of the capital stock:

$$Dk = \left(\frac{g_{11}'' + g_{22}'' - 2g_{12}''}{\Delta}\right)\rho^*\theta. \tag{A8}$$

Taking the ratio between (A7) and (A8),

$$\frac{dk}{Dk} = \frac{g_{22}'' - g_{12}''}{g_{11}'' + g_{22}'' - 2g_{12}''}.$$

Since the production function $g(k_1, k_2)$ is homogeneous, it can be rewritten under the form $g(k_1, k_2) = H[F(k_1, k_2)]$, where F is homogeneous of degree one.

Use the properties of F, and the equality of F_1' and F_2' (at $\theta = 0$), to write

$$g''_{22} - g''_{12} = H'(F''_{22} - F''_{12}) = H'F''_{22}\left(1 + \frac{k_2}{k_1}\right),$$

$$g''_{11} + g''_{22} - 2g''_{12} = H'F''_{22}\left[1 + 2\frac{k_2}{k_1} + \left(\frac{k_2}{k_1}\right)^2\right].$$

Taking the ratio between these relations, $dk/Dk = k_1/k$. This result could be generalized easily. The long-run incidence of a sectoral capital income tax on the total capital stock is proportional to the share of the taxed capital with respect to the total capital stock.

References

Arrow, Kenneth J., and Kurz, Mordecai. *Public Investment, the Rate of Return and Optimal Fiscal Policy.* Baltimore: Johns Hopkins Univ. Press, 1970.

Barro, Robert J. "Are Government Bonds Net Wealth?" *J.P.E.* 82, no. 6 (November/December 1974): 1095–1117.

Boadway, Robin, and Treddenick, John. "The Effects of the U.S. Corporate Tax on Resource Allocation and Welfare." Mimeographed. Kingston, Canada: Queen's Univ., March 1975.

Chamley, Christophe. "The Stability of the Capital Accumulation in a Competitive Economy with Imperfect Foresight." Mimeographed. New Haven, Conn.: Yale Univ., December 1979.

———. "Optimal Intertemporal Taxation and the Public Debt." Mimeographed. New Haven, Conn.: Yale Univ., April 1980. (*a*)

———. "Optimal Taxation in a Life-Cycle Capital Accumulation Model." Mimeographed. New Haven, Conn.: Yale Univ., May 1980. (*b*)

Diamond, Peter A. "National Debt in a Neoclassical Growth Model." *A.E.R.* 55 (December 1965): 1126–50.

———. "Incidence of an Interest Income Tax." *J. Econ. Theory* 2 (September 1970): 211–24.

Ebrill, Liam P., and Hartman, David G. "On the Incidence and Excess Burden of the Corporation Tax." Mimeographed. Cambridge, Mass.: Harvard Univ., October 1977.

Feldstein, Martin S. "Incidence of a Capital Income Tax in a Growing Economy with Variable Savings Rates." *Rev. Econ. Studies* 41 (October 1974): 505–13. (*a*)

———. "Tax Incidence in a Growing Economy with Variable Factor Supply." *Q.J.E.* 88 (November 1974): 551–73. (*b*)

———. "The Welfare Cost of Capital Income Taxation." *J.P.E.* 86, no. 2, suppl. (April 1978): S29–S51.

Friedländer, Ann F., and Vandendorpe, Adolf F. "Capital Taxation in a Dynamic General Equilibrium Setting." *J. Public Econ.* 10 (August 1978): 1–24.

Fullerton, Don; Shoven, John; and Whalley, John. "Dynamic General Equilibrium Impacts of Replacing the U.S. Income Tax with a Progressive Consumption Tax." Paper presented at the conference on the Taxation of Capital, Nat. Bur. Econ. Res., Cambridge, Mass., November 1979.

Green, Jerry R., and Sheshinski, Eytan. "Approximating the Efficiency Gain of Tax Reforms." *J. Public Econ.* 11 (April 1979): 179–195.

Hall, Robert E. "Consumption Taxes versus Income Taxes: Implications for Economic Growth." Paper presented at the 61st Annual Conference on Taxation, 1969.

———. "The Dynamic Effects of Fiscal Policy in an Economy with Foresight." *Rev. Econ. Studies* 38 (April 1971): 229–44.

Harberger, Arnold C. "The Incidence of the Corporation Income Tax." *J.P.E.* 70, no. 3 (June 1962): 215–40.

Harberger, Arnold C., and Bruce, Neil. "The Incidence and Efficiency Effects of Taxes on Income from Capital: A Reply." *J.P.E.* 84, no. 6 (December 1976): 1285–92.

Hudson, Edwards, and Jorgenson, Dale. "The Tax on Capital Income in the U.S." Report to Office of Taxation, U.S. Treasury Dept., Washington, January 1976.

Iwai, Katsuhito. "Optimal Economic Growth and Stationary Ordinal Utility—a Fisherian Approach." *J. Econ. Theory* 5 (August 1972): 121–51.

Jorgenson, Dale W., and Christensen, Laurits R. "Measuring Economic Performance in the Private Sector." In *The Measurement of Economic and Social Performance.* Studies in Income and Wealth, vol. 38. Edited by Milton Moss. New York: Columbia Univ. Press (for Nat. Bur. Econ. Res.), 1973.

Koopmans, Tjalling C. "Objectives, Constraints and Outcomes in Optimal Growth Models." *Econometrica* 35 (January 1967): 1–15.

Krzyzaniak, Marian. "The Long-Run Burden of a General Tax on Profits in a Neoclassical World." *Public Finance* 22, no. 4 (1967): 472–91.

Laitner, John. "Household Bequests, Perfect Expectations, and the National Distribution of Wealth." *Econometrica* 47 (September 1979): 1175–93.

Levhari, David, and Sheshinski, Eytan. "Lifetime Excess Burden of a Tax." *J.P.E.* 80, no. 1 (January/February 1972): 139–47.

Sato, Kazuo. "On the Adjustment Time in Neo-classical Growth Models." *Rev. Econ. Studies* 33 (July 1966): 263–68.

———. "Taxation and Neo-classical Growth." *Public Finance* 22, no. 3 (1967): 346–70.

Shoven, John B. "The Incidence and Efficiency Effects of Taxes on Income from Capital." *J.P.E.* 84, no. 6 (December 1976): 1261–83.

Shoven, John B., and Whalley, John. "A General Equilibrium Calculation of the Effects of Differential Taxation of Income from Capital in the U.S." *J. Public Econ.* 1 (November 1972): 281–321.

Summers, Lawrence H. "Tax Policy in a Life-Cycle Model." Working Paper no. 302, Nat. Bur. Econ. Res., Cambridge, Mass., November 1979.

Weber, Warren E. "Interest Rates, Inflation, and Consumer Expenditures." *A.E.R.* 65 (December 1975): 843–58.

Wright, Colin. "Saving and the Rate of Interest." In *The Taxation of Income from Capital,* edited by Arnold C. Harberger and Martin J. Bailey. Washington: Brookings Inst., 1969.

Foreign Ownership and the Theory of Trade and Welfare

Richard A. Brecher

Carleton University

Jagdish N. Bhagwati

Columbia University

Some standard topics in the theory of international trade are reconsidered in this paper by distinguishing between national and aggregate income when fixed supplies of foreign inputs are present within the home country. Under conditions that would ensure a national welfare gain if foreign ownership were absent, international transfer, economic growth, or tariff policy might cause a national welfare loss in the presence of foreign ownership. The techniques developed could be applied to other domestic distinctions (such as those based on race, sex, age, or ethnicity) and to the theory of customs unions in a three-country world.

I. Introduction

This paper reconsiders a number of standard topics in the theory of international trade by taking explicit account of the distinction between national and aggregate income when fixed supplies of foreign-owned inputs are present within the domestic economy. Extending the work of Bhagwati and Brecher (1980),[1] the following

Thanks are due to the National Science Foundation grant no. SOC79-07541 for partial financial support of the research underlying this paper. The comments and suggestions of Alan Deardorff, Jacob Frenkel, Alasdair Smith, and anonymous referees are gratefully acknowledged.

[1] Their work, in turn, extends the analysis of Bhagwati and Tirom (1980), who concentrate on a special case mentioned in n. 3 below.

[*Journal of Political Economy*, 1981, vol. 89, no. 3]
© 1981 by The University of Chicago. 0022-3808/81/8903-0004$01.50

analysis takes a new look at welfare-theoretic aspects of international transfer, economic expansion, and tariff policy, while it emphasizes significant departures from conventional wisdom that arise in the presence of foreign ownership. As these selected departures suggest, many standard results are open to serious question when part of the domestic product accrues to factor inputs from abroad.

Originally, the motivation for the present two-group analysis (based on the national-foreign distinction) came from a recent concern in Latin America, where policymakers have been worried about the impact of trade liberalization on national welfare, given the domestic presence of foreign-owned multinational corporations. After further reflection, however, it is clear that the treatment below has much greater applicability to a broad range of analytically similar cases. For example, it is possible to treat in much the same way a wide variety of alternative domestic distinctions, including those based on race, sex, age, or ethnicity. The following techniques and results, moreover, are directly relevant for the fully analogous two-group issue relating to the distribution of gains (or losses) between trading partners in a customs union (such as the European Economic Community) with factor mobility. While these other policy problems are of considerable importance and interest as well, only the national-foreign distinction is pursued explicitly here for the sake of brevity.

Section II reviews the basic model of an open economy, in which foreign-owned and national supplies of two homogeneous factors are combined to produce two commodities. As Section III then shows, a transfer-receiving country might suffer a loss in national welfare, even under the usual conditions which would ensure a welfare gain if foreign ownership were absent. As established next by Section IV, a country experiencing economic expansion (due to factor-endowment growth or technological advance) might encounter a deterioration in national welfare, even under well-known conditions which would preclude this possibility of "immiserizing growth" in the absence of foreign ownership. Afterward, Section V explains why free trade might be inferior to *both* no trade and subsidized trade, as far as national welfare is concerned.[2] Section VI summarizes the paper's main results, based on the possibility of aggregate and national welfare moving in opposite directions.

Needless to say, this possibility would not arise if foreign-owned factors were taxed to the nationally optimal extent. Indeed, with these factors in perfectly inelastic supply, the optimal tax on each foreign

[2] This result is obtained also by Bhagwati and Tironi (1980), for a special case identified in n. 3 below. In addition, since Bhagwati and Brecher (1980) compare free-trade equilibrium with autarky, the present paper will emphasize instead the comparison of free versus subsidized trade.

input clearly would be 100 percent, thereby removing the after-tax distinction between aggregate and national welfare. Assuming that this type of optimal taxation of factors is politically infeasible, however, the present analysis cautions nationally oriented policymakers against the usual, automatic adoption of the standard welfare conclusions which reflect an aggregate point of view. More specifically, this paper shows precisely how the traditional (aggregate) propositions must be modified for a truly national perspective, when political constraints eliminate optimal taxation of inputs from abroad.

II. The Basic Model

Following the analysis of Bhagwati and Brecher (1980), the present section summarizes the basic two-commodity, two-factor model of an open economy (large or small), which plays host to given quantities of inputs from abroad. The aggregate factor endowments of the country are fixed at $\bar{K}^a$ units of capital and $\bar{L}^a$ units of labor, while the given amounts $\bar{K}^n$ and $\bar{L}^n$ are the national endowments of capital and labor, respectively. (Thus, the fixed supplies of foreign-owned capital and labor within the home country are $\bar{K}^a - \bar{K}^n$ and $\bar{L}^a - \bar{L}^n$, respectively.) It is assumed that $\bar{K}^a > \bar{K}^n > 0$ and $\bar{L}^a > \bar{L}^n > 0$, excluding the possibility that either factor within the home economy is owned wholly by nationals or completely by foreigners.[3] Commodity two is always labor intensive relative to capital-intensive commodity one, and the well-behaved technology exhibits constant returns to scale.

In figure 1, the home country is depicted in free-trade equilibrium. Aggregate production is at point Q^a on production-possibility frontier $T_2^aT_1^a$ (corresponding to $\bar{K}^a$ and $\bar{L}^a$), aggregate income is represented by budget line Q^aD^a, and aggregate consumption occurs at point D^a on indifference curve $I_2^aI_1^a$. (For simplicity of exposition, it is assumed that all income earned by factors from abroad is consumed locally, to avoid having to show repatriation of such income within the diagram.) By the reasoning of Bhagwati and Brecher (1980), national consumption takes place at point D^n on indifference curve $I_2^nI_1^n$, with national income given by budget line Q^nD^n (parallel to Q^aD^a), *as if* nationals produced separately at point Q^n on production-possibility frontier $T_2^nT_1^n$ (drawn for $\bar{K}^n$ and $\bar{L}^n$).[4] To emphasize that the main results of this paper qualitatively do *not* require any differences in consumer preferences between nationals and foreigners within the home country, assume throughout the text that the same set of indifference

[3] For the special case in which $\bar{K}^a > \bar{K}^n = 0$ and $\bar{L}^a = \bar{L}^n > 0$, see Bhagwati and Tironi (1980).

[4] The discussion could be extended readily to allow for the possibility of complete specialization, following the analysis of Bhagwati and Brecher (1980)

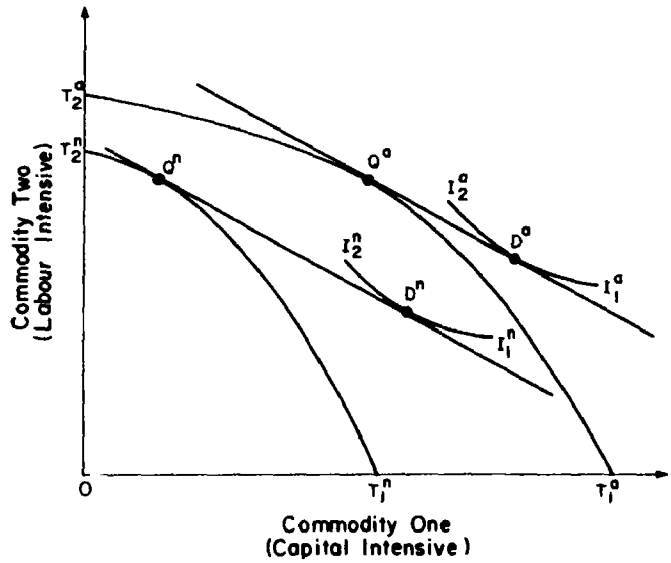


FIG. 1.—Differential trade-volume phenomenon

curves with unitary income elasticities of demand represents both national and aggregate tastes in consumption, although this simplification of the exposition could be dropped (as in footnotes to this paper) without detracting from the essence of the analysis.[5]

The model may be summarized conveniently as follows:

$$X_i^j = F_i^j(p), i = 1, 2, j = a, n; \tag{1}$$

$$Y^j = X_1^j + pX_2^j, j = a, n; \tag{2}$$

$$W^j = U^j(C_1^j, C_2^j), j = a, n; \tag{3}$$

$$C_1^j + pC_2^j = Y^j, j = a, n; \tag{4}$$

where p denotes the relative price of the second commodity in terms of the first; X_i^j denotes output of commodity i on frontier $T_2^jT_1^j$; each F_i^j is a conventional function of p, given $\bar{K}^j$, $\bar{L}^j$ and the (uniform) technology for commodity i; Y^a and Y^n denote the real value of aggregate and national income, respectively, in terms of the first commodity; C_i^a and C_i^n denote aggregate and national consumption, respectively, of commodity i ($i = 1, 2$); W^a and W^n denote aggregate and national welfare,

[5] Nn. 6–8, 10, and 15 below extend the discussion to let tastes differ between nationals and foreigners within the home country. These extensions bring out the essentially "three-country" flavor of the analysis, in which nationals, domestically located foreigners, and the rest of the world can be treated as three distinct components of the international economy.

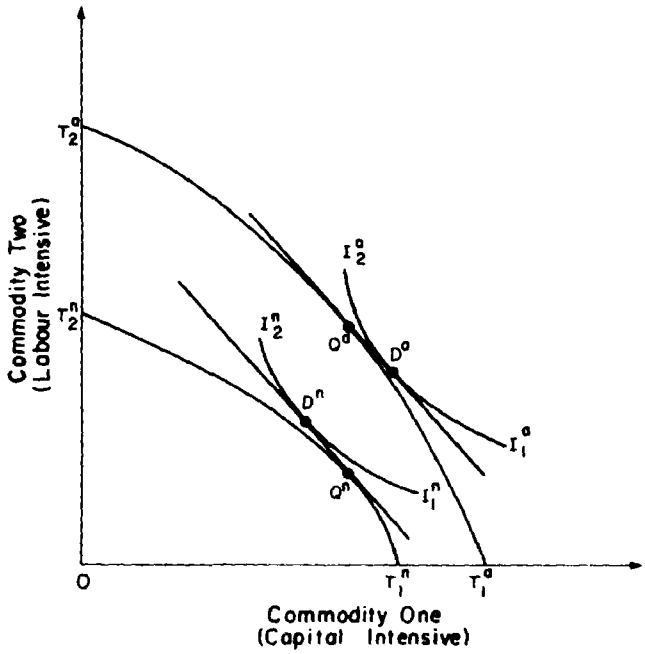


FIG. 2.—Differential trade-pattern phenomenon

respectively; and each U^j is a concave function of C_1^j and C_2^j, with positive partial derivatives denoted by $U_i^j \equiv \partial U^j/\partial C_i^j$ $(i = 1, 2)$.

Later in the paper when a change in relative prices is induced by various parametric shifts, either of the following phenomena might lead to a fall in national despite a rise in aggregate welfare, depending on the strength of other induced effects. The differential trade-volume phenomenon is shown in figure 1, where the aggregate (actual) volume of trade (defined by line segment Q^aD^a) is less than the national (hypothetical) volume of trade (defined by line segment Q^nD^n), implying (ceteris paribus) that a terms-of-trade deterioration worsens national by more than aggregate welfare. Figure 2 (labeled similarly) illustrates the Bhagwati and Brecher (1980) differential trade-pattern phenomenon, which arises when the aggregate and national patterns of trade differ (in direction), so that an aggregate terms-of-trade improvement (tending to raise W^a) means a national terms-of-trade deterioration (tending to lower W^n). The national relative to the aggregate endowment of factors is labor abundant in figure 1 (with $\bar{K}^a/\bar{L}^a > \bar{K}^n/\bar{L}^n$) but capital abundant in figure 2 (with $\bar{K}^a/\bar{L}^a < \bar{K}^n/\bar{L}^n$), as suggested by the relative shapes of frontiers $T_2^aT_1^a$ and $T_2^nT_1^n$, in accordance with the reasoning of Rybczynski (1955).

To understand the possibility of a fall in national welfare despite a rise in aggregate welfare, it might be tempting to go no further than the following simple observation. Whenever the national and aggregate endowments exhibit different capital/labor ratios, the domestic distribution of income might deteriorate for nationals, as a change in relative commodity prices alters the wage/rental ratio for reasons expounded by Stolper and Samuelson (1941). It is important to recognize, however, that generally this income-redistribution effect will not be strong enough to produce the differential responses in national and aggregate welfare if the relative factor-endowment discrepancy is too small to create either the differential trade-volume or the differential trade-pattern phenomenon. Even when either of these phenomena arises, moreover, a fall in national welfare despite a rise in aggregate welfare can occur if and only if certain specific conditions (derived below) are satisfied.

III. International Transfer

According to a standard result in the literature (see Mundell 1960), a transfer-receiving country cannot suffer a loss in aggregate welfare despite any possible deterioration in the aggregate terms of trade, as long as international commodity-market equilibrium is stable. In other words, the transfer-induced change in W^a cannot be negative, assuming that an excess demand for or supply of the second good in world markets can be cleared by a rise or fall in p, respectively. As the following argument demonstrates, however, a (large) transfer-receiving country might suffer a deterioration in national welfare, even under the assumption (maintained throughout the present paper) that commodity markets are stable. This demonstration of a transfer-induced fall in W^n, moreover, does not even require a rise in the relative price of home importables.

If it is assumed that the transfer is given only to nationals, equations (2) are modified as follows:

$$Y^j = X_1^j + pX_2^j + \tau, j = a, n, \tag{5}$$

where τ is the real value of the transfer in terms of the first commodity. If any part of the transfer were given to foreigners within the home country, the chances for a decline in W^n would simply be enhanced, thereby strengthening the argument below.

To examine the welfare implications of the transfer, differentiate equations (1), (3), (4), and (5) totally with respect to τ—assuming (without loss of generality) that initially $U_1^j = 1$, while noting that $U_2^j/U_1^j = p = -(dF_1^j/dp)/(dF_2^j/dp)$ from the first-order conditions for

maximizing utility and profit. In this way, it is a straightforward exercise to derive

$$dW^j/d\tau = 1 + (E^j dp/d\tau), j = a, n, \tag{6}$$

where $E^j = X_2^j - C_2^j$. Consistent with figures 1 and 2, which depict the home country exporting the second good, $E^a > 0$ by assumption throughout the present paper. As illustrated above, however, E^n can be either positive (in fig. 1) or negative (in fig. 2).

As equations (6) confirm, $dW^a/d\tau$ is the familiar sum of the following two components: the primary gain (= 1) from the transfer-induced increase in aggregate income, at the initial (pretransfer) set of relative prices; plus the secondary effect (= $E^a dp/d\tau$) from the possible increase or decrease in the real exchange value of the initial volume of home exports, in the event of a transfer-induced change (if any) in relative prices. The expression for $dW^n/d\tau$ is analogous. If foreign inputs were entirely absent from the home country, the distinction between national and aggregate variables would disappear, thereby implying that $E^n = E^a$ and (hence) that $dW^n/d\tau = dW^a/d\tau$. Given the actual presence of factor inputs from abroad, however, $dW^n/d\tau$ generally differs from $dW^a/d\tau$, except in the special case where either $E^n = E^a$ (despite the foreign presence) or $dp/d\tau = 0$.

To determine precise conditions for the direction of change in welfare, consider the standard transfer-induced terms-of-trade response, analyzed previously by Samuelson (1952, 1954) and subsequently by Mundell (1960). Thus, by well-known reasoning,

$$dp/d\tau = (1 - m - m^*)/(e + e^* - 1)E^a, \tag{7}$$

where e (> 0) and m denote the relative-price elasticity of import demand and the marginal propensity to consume the importable, respectively, for the home country; e^* (> 0) and m^* denote the corresponding variables for the rest of the world; and $\tau = 0$ in the initial (pretransfer) equilibrium.[6] Given the assumption above that world commodity-market equilibrium is stable, $e + e^* > 1$ throughout the present paper.

If equation (7) is substituted into equations (6), simple manipulation confirms that

$$dW^a/d\tau = (\epsilon + \epsilon^*)/(e + e^* - 1) > 0, \tag{8}$$

[6] If preferences in consumption were allowed to differ between nationals and foreigners within the home country, it would be necessary to rewrite eq. (7) as follows, to reflect the present assumption that the entire transfer goes exclusively to nationals:

$$dp/d\tau = (1 - m^n - m^*)/(e + e^* - 1)E^a, \tag{7'}$$

where m^n denotes the national marginal propensity to consume the home importable.

but shows that

$$dW^n/d\tau \lesseqgtr 0 \text{ as } (e + e^* - 1)E^a \lesseqgtr (m + m^* - 1)E^n, \tag{9}$$

where ϵ (> 0) and ϵ^* (> 0) denote the compensated (constant-utility) relative-price elasticity of import demand for the home country and the rest of the world, respectively, while $e = \epsilon + m$ and $e^* = \epsilon^* + m^*$, according to a standard decomposition.[7] Although $dW^a/d\tau > 0$ unambiguously, it is evidently possible to have $dW^n/d\tau < 0$ nevertheless.[8]

To highlight the important role of the differential trade-pattern and differential trade-volume phenomena, it is helpful to revert to equations (6), which imply that a fall in national despite the rise in aggregate welfare can occur only if $(E^n - E^a)dp/d\tau < 0$. This necessary condition for a fall in W^n holds if either $dp/d\tau < 0$ in presence of the differential trade-volume phenomenon of figure 1 (where $E^n > E^a > 0$) or $dp/d\tau > 0$ in conjunction with the differential trade-pattern phenomenon of figure 2 (where $E^n < 0 < E^a$).[9] Correspondingly, if home exportables were relatively intensive in their use of capital (rather than labor), a transfer-induced deterioration in national

[7] Alternatively, if eq (7′) from n. 6 above were substituted into eqq. (6), simple manipulation could show that

$$dW^a/d\tau = [(\epsilon + \epsilon^*) + (1 - \gamma)(m^f - m^n)]/(e + e^* - 1) \tag{8′}$$

and

$$dW^n/d\tau = [(\epsilon + \epsilon^*) + (1 - \gamma)(m^f + m^* - 1)]/(e + e^* - 1), \tag{9′}$$

where m^f denotes the marginal propensity to consume the home importable for foreigners within the home country; $\gamma \equiv (C_1^n - X_1^n)/(C_1^a - X_1^a) = E^n/E^a$; and use is made of the fact that $m = \gamma m^n + (1 - \gamma)m^f$. Eq. (8′) indicates that $dW^a/d\tau$ can be decomposed into two comparative-static components. As could be shown readily, the first component $[(\epsilon + \epsilon^*)/(e + e^* - 1)]$ is the transfer-induced change in W^a that would occur initially if the transfer were given temporarily to nationals and domestically located foreigners in the respective amounts $\gamma\tau$ and $(1 - \gamma)\tau$, whereas the second component $[(1 - \gamma)(m^f - m^n)/(e + e^* - 1)]$ is the subsequent change in W^a that would occur as the portion $(1 - \gamma)\tau$ was passed from domestically located foreigners to nationals (the ultimate recipients of the entire transfer). Eq. (9′) could be interpreted analogously, since $-dW^n/d\tau$ equals the worldwide sum of transfer-induced changes in welfare for everyone excluding home-country nationals, as could be shown readily.

[8] Under the present assumption that $m^f = m^n = m$, eq. (8′) of n. 7 above is equivalent to eq. (8), while eq. (9′) leads directly to condition (9). Alternatively, if it were the case that $m^f \neq m^n$, there would arise the new possibility of having $dW^a/d\tau < 0$ in eq. (8′). Also if it were supposed that $m^f = 1 - m^*$, it would be the case that $dW^n/d\tau > 0$ unambiguously in eq. (9′). This last result can be understood intuitively as follows: If foreign tastes are uniform throughout the world, the reasoning behind eq. (8) shows equally well that the transfer must lower worldwide foreign welfare; that is, $-dW^n/d\tau < 0$, recalling n. 7 above. Incidentally, in view of the fact that worldwide foreign welfare otherwise can rise (when $dW^n/d\tau < 0$) if $m^f \neq m^*$, international aid might be especially attractive for a donor country with investments in the aid-receiving economy.

[9] Although $(E^n - E^a)dp/d\tau < 0$ also if $0 < E^n < E^a$ when $dp/d\tau > 0$, $dW^n/d\tau > 0$ in this case, as implied by eqq. (6). The reader may also see alternatively that, from condition (9), $dW^n/d\tau < 0$ if and only if $(\epsilon + \epsilon^*)E^n + (E^a - E^n)(e + e^* - 1) < 0$. Therefore, national welfare *may* decline despite the increase in aggregate welfare if $E^n < 0 < E^a$ (i.e., the differential trade-pattern phenomenon holds) or if $E^n > E^a > 0$ (i.e., the differential trade-volume phenomenon holds).

(though not in aggregate) welfare would still be possible, provided that either the aggregate terms of trade improve in the case of labor-abundant nationals or an aggregate terms-of-trade decline occurs in the presence of capital-abundant nationals.

Consequently, the basic results of this section can be summarized generally in the following terms. When the home exportable uses intensively the factor that is relatively abundant in the national (as compared with the aggregate) endowment, the national and aggregate patterns of trade are the same, in which case a fall in national welfare might occur through a differential trade-volume phenomenon if the (national and aggregate) terms of trade worsen unambiguously. Alternatively, when the home exportable uses intensively the factor that is relatively scarce in the national (as compared with the aggregate) endowment, the aggregate and national patterns of trade could differ, in which case the differential trade-pattern phenomenon might give rise to a deterioration in national welfare if the national terms of trade worsen through an aggregate terms-of-trade improvement. These general results, moreover, hold equally well for changes in p induced by economic expansion and tariff policy, as will be clear from the analysis below.

IV. Economic Expansion

As Bhagwati (1958*a*) has demonstrated, a once-for-all increase in a factor endowment or in a technological level might deteriorate the aggregate terms of trade enough to worsen aggregate welfare of the home country, but this immiserizing growth can occur only if either the rest of the world has an inelastic offer curve or growth would decrease the production of home importables at the initial product-price ratio. In other words, if the offer-curve elasticity for the rest of the world is not less than unitary and economic expansion is not "ultrabiased" against the production of home importables, the growth-induced change in W^a cannot be negative. Even under these circumstances (assumed throughout the present section) which preclude a fall in aggregate welfare, however, the following analysis demonstrates that a (large) country might suffer a loss in national welfare. This demonstration of a growth-induced decline in W^n, moreover, does not even require a rise in the relative price of home importables.

To allow for factor-endowment expansion or technological advance, equations (1) may be rewritten as follows:

$$X^j_i = F^j_i(p, \theta), i = 1, 2, j = a, n, \tag{10}$$

where θ is a general shift parameter, a rise in which indicates either a factor-endowment increase (for $\bar{K}^j$ or $\bar{L}^j$) or a disembodied

technological improvement for an industry (one or two). It is assumed that any addition to the aggregate supply of capital or labor is owned fully by nationals. If any part of such addition were foreign owned, the likelihood of a decline in W^n would simply be enhanced, thereby strengthening the argument below. However, the ability of domestically located producers to take advantage of disembodied technological progress should be independent of the source of ownership of the inputs used, as assumed here.

Differentiating equations (2), (3), (4), and (10) totally with respect to θ, while recalling that $U^j_1 = 1$ initially and that $U^j_2/U^j_1 = p = -(\partial F^j_1/\partial p)/(\partial F^j_2/\partial p)$, we readily obtain the following result:

$$dW^j/d\theta = Y^j_\theta + (E^j dp/d\theta), j = a, n, \tag{11}$$

where $Y^j_\theta \equiv \partial Y^j/\partial\theta > 0$. Thus, each $dW^j/d\theta$ is the sum of a primary growth effect (Y^j_θ) plus a secondary relative-price effect ($E^j dp/d\theta$), which are analogous to the welfare-related effects of the transfer mentioned above in Section III. Although national and aggregate welfare again would remain equal if foreign inputs were entirely absent from the home country, the actual presence of foreign ownership gives rise to the possibility of having $dW^n/d\theta < 0$ when $dW^a/d\theta > 0$, except in the special case where $dp/d\theta = 0$.

Turning to the standard growth-induced terms-of-trade response, analyzed previously by Bhagwati (1958*b*) and subsequently by Kemp (1969, p. 110), we see that it is a well-known fact that

$$dp/d\theta = (\beta - m)Y^a_\theta/(e + e^* - 1)E^a, \tag{12}$$

where $\beta \equiv (\partial X^a_1/\partial\theta)/Y^a_\theta$. When this result is substituted into equations (11), straightforward manipulation confirms that

$$dW^a/d\theta = (\epsilon + \beta + e^* - 1)Y^a_\theta/(e + e^* - 1) > 0 \tag{13}$$

but shows that

$$dW^n/d\theta \gtreqless 0 \text{ as } E^a(e + e^* - 1)Y^a_\theta \gtreqless E^n(m - \beta)Y^a_\theta, \tag{14}$$

where $\beta \geqq 0$, which recalls the assumption that growth would not reduce production of home importables at the initial commodity-price ratio; and $e^* \geqq 1$, which recalls the assumption that the rest of the world's offer curve is not inelastic. Thus, despite the fact that $dW^a/d\theta > 0$ unambiguously under these circumstances, it is still possible to have $dW^n/d\theta < 0$ nevertheless.[10]

[10] Along lines suggested by nn. 6–8 for the case of international transfer, the analysis of economic expansion could be extended readily to distinguish between m^n and m^f. It is worth noting, however, that it would still be possible to have $dW^n/d\theta < 0$ even if it were the case that $m^f = 1 - m^*$.

Equations (11) imply that a fall in national despite the rise in aggregate welfare can occur only if $(E^n - E^a)dp/d\theta < Y_2^a - Y_2^n$. As could be shown readily, this necessary condition for a fall in W^n can result from each of the following alternative events, for example: an increase in either the national endowment of capital or the technological level of industry one, with $dp/d\theta > 0$ in the presence of the differential trade-pattern phenomenon (fig. 2); and an increase in either the national endowment of both factors or the technological level of both industries, if $dp/d\theta < 0$ with the differential trade-volume phenomenon (fig. 1).[11] Correspondingly, if home exportables were relatively capital intensive, it would be possible to have an expansion-induced deterioration in national (though not in aggregate) welfare under a variety of circumstances, including the following: an increase either in the national stock of labor or in the level of technology for the production of importables, when the national endowment is labor abundant; or an increase either in the national endowment of both factors or in the level of technology for both sectors, when nationals are capital abundant.

V. Tariff Policy

According to a standard result in the literature (see Bhagwati 1968), free trade is ranked superior to both no trade and subsidized trade (assuming that both offer curves are well behaved),[12] from the viewpoint of aggregate welfare. In other words, the home country cannot increase W^a above the free-trade level either by using an import (or export) tax to eliminate trade or by imposing an export (or import) subsidy to encourage trade. From the national-welfare point of view, however, the ranking above may be reversed. Since Bhagwati and Brecher (1980) already demonstrated the possibility of such a reversal for free trade versus autarky, the following analysis concentrates on free versus subsidized trade.

To allow for tariff policy, equations (2) may be modified as follows:

$$Y^j = X_1^j + pX_2^j + [(C_1^a - X_1^a)\,\alpha/(1 - \alpha)],\ j = a, n; \qquad (15)$$

where α denotes the *ad valorem* tariff, which is an import tax (if $\alpha > 0$) or an import subsidy (if $\alpha < 0$); the domestic relative price of the

[11] Although $Y_2^a = Y_2^n$ with a national factor-endowment increase, it is possible that $Y_2^a > Y_2^n$ for a technological advance. Thus, with the latter (but not the former) type of economic expansion, a fall in national welfare despite the rise in aggregate welfare might occur even without the differential trade-volume and differential trade-pattern phenomena—if both industries experience the technological advance and $dp/d\theta < 0$.

[12] I.e., the offer curve is assumed to represent imports as a monotonic decreasing function of their relative price in world markets. For the significance of this assumption in tariff analysis, see Bhagwati and Kemp (1969).

second good is still denoted by p, so that the relative price of this good in world markets is now equal to $p(1 + \alpha)$; and $(C_1^a - X_1^a)\alpha/(1 + \alpha)$ equals the real value (in terms of the first good) of tax revenues or subsidy payments, evaluated at domestic prices.[13] In writing equations (15), it is assumed (for the sake of simplicity) that all tax revenues or subsidy payments, respectively, are returned to or collected from *national* consumers in lump-sum fashion. If foreigners within the home country were to receive or finance any part of these revenues or payments, respectively, comparison of the free-trade and autarkic equilibria (which generate no tax revenues) clearly would be unaffected, while the chances of having free trade inferior to subsidized trade simply would be enhanced (thereby strengthening the analysis below).

To show that free trade might be inferior to subsidized trade from the national point of view, it is sufficient to establish the possibility of having $dW^n/d\alpha < 0$ in free-trade equilibrium. Consequently, throughout the following discussion, let $\alpha = 0$ in the initial (pretariff) equilibrium.

Differentiating equations (1), (3), (4), and (15) totally with respect to α and again recalling that $U_1^j = 1$ initially and that $U_2^j/U_1^j = p = -(dF_1^j/dp)/(dF_2^j/dp)$, we may verify readily that

$$dW^j/d\alpha = pE^a + (E^j dp/d\alpha), j = a, n; \tag{16}$$

note that $pE^a = C_1^a - X_1^a$ when (balanced) trade is initially free (with $\alpha = 0$). By well-known reasoning (see Kemp 1969, p. 96),

$$dp/d\alpha = p(1 - m - e^*)/(e + e^* - 1); \tag{17}$$

note that $1 - m$ equals the home country's marginal propensity to consume the exportable and recall that $\alpha = 0$ initially. When this result is substituted into equations (16), simple manipulation confirms that[14]

$$dW^a/d\alpha = \epsilon pE^a/(e + e^* - 1) \geqq 0 \tag{18}$$

but shows that

$$dW^n/d\alpha \gtreqless 0 \text{ as } (e + e^* - 1)E^a \lesseqgtr (m + e^* - 1)E^n. \tag{19}$$

Thus, despite the fact that $dW^a/d\alpha \geqq 0$, it is evidently possible to have $dW^n/d\alpha < 0$ nevertheless.[15]

[13] Although the corresponding value at *world* prices would be $(C_1^a - X_1^a)\alpha$, consumers respond directly to *domestic* (tariff-inclusive) prices instead.

[14] Note that $dW^a/d\alpha = 0$ only in the small-country case where $e^* = \infty$, and even then the change in W^a does not equal zero for any discrete change in α, by well-known reasoning.

[15] Along lines suggested by nn. 6–8, the analysis of tariff policy could be extended readily to let $m^f \neq m^n$, without eliminating the possibility of having $dW^n/d\alpha < 0$ even if $m^f = 1 - m^*$.

As implied by the equations (16), it is possible to have $dW^n/d\alpha < 0$ (even though $dW^a/d\alpha$ cannot be negative) if either a differential trade-volume phenomenon arises (fig. 1) when $dp/d\alpha < 0$ (the "normal" price response) or a differential trade-pattern phenomenon occurs (fig. 2) when $dp/d\alpha > 0$. (The "perverse" price response [$dp/d\alpha > 0$] can occur only in the large-country case, under conditions discussed by Metzler [1949].) Correspondingly, if home exportables were relatively capital intensive, it would be possible to have $dW^n/d\alpha < 0$ (even though $dW^a/d\alpha$ cannot be negative) if either nationals are labor abundant when $dp/d\alpha > 0$ or nationals are capital abundant when $dp/d\alpha < 0$.

Thus a trade subsidy might raise national (but not aggregate) welfare above the free-trade level. This analysis of a small subsidy (tax) on trade, moreover, complements the discussion of Bhagwati and Brecher (1980), who concentrate on prohibitive taxes on trade and thus are able to avoid the issue of tariff revenues.

The analysis of this section has an important implication for the traditional method of estimating the cost (benefit) of tariff protection or trade liberalization. Since the conventional method (as outlined by Johnson [1960]) ignores the source of ownership of domestically located inputs, the concept measured (in present notation) is clearly $dW^a/d\alpha$ rather than $dW^n/d\alpha$. Thus, the traditional estimate of the impact of protection or liberalization is an aggregate measure, which overstates or understates the national cost (benefit) if $(E^n - E^a)dp/d\alpha \lesseqgtr 0$, respectively, as suggested by equations (16). This misstatement arises because the conventional estimate simply sums the three standard components (namely, the external terms-of-trade effect and the costs of distortion in both production and consumption), while it fails to exclude the foreign-factor portion of the tariff-induced change in aggregate welfare.[16]

VI. Summary

As demonstrated by this paper, welfare aspects of international trade theory need to be reconsidered, when national and aggregate income

[16] The foreign-factor portion of the change in aggregate welfare is represented by the expression $(E^a - E^n)dp/d\alpha$, which must be excluded from $dW^a/d\alpha$ to give $dW^n/d\alpha$, as suggested by eqq. (16). Also by repeating the procedure of Bhagwati, Ramaswami, and Srinivasan (1969), it is possible to write (in present notation) that $dW^a/d\alpha = (E^a dp^*/d\alpha) + [(p^* - p)dX_2^a/d\alpha] + [(p - p^*)dC_2^a/d\alpha]$, where the relative price of the second good in world markets is denoted p^*, which equals $p(1 + \alpha)$. The components $E^a dp^*/d\alpha$, $(p^* - p)dX_2^a/d\alpha$, and $(p - p^*)dC_2^a/d\alpha$ are the effects due to the terms-of-trade change, the production distortion, and the consumption distortion, respectively. When evaluated in free-trade equilibrium (where $\alpha = 0$ and $p = p^*$), the latter two (distortion-related) components disappear, leaving only the first (terms-of-trade) component. This remaining (first) component, moreover, is equivalent to the right-hand side of eqq. (16) for $j = a$, since (in free-trade equilibrium) $dp/d\alpha = (dp^*/d\alpha) - p$.

differ in the presence of foreign ownership. Examples of this need are provided by the analysis of international transfer, economic expansion, and tariff policy. For a country receiving a transfer from abroad, national (but not aggregate) welfare might deteriorate even when international commodity-market equilibrium is stable, regardless of the direction of change in the world product-price ratio. In the case of economic expansion from factor-supply growth or technological advance, national (but not aggregate) welfare might worsen even when the rest of the world does not have an inelastic offer curve and domestic expansion is not ultrabiased against production of home importables, no matter what the direction of change in the world commodity-price ratio. As for tariff policy, free trade might be ranked inferior to both no trade and subsidized trade (in either direction), from the viewpoint of national (but not aggregate) welfare. Moreover, the conventional empirical estimates of the cost of protection à la Johnson's (1960) methodology are generally seen to be in need of correction if the economy has foreign-owned factors of production. In fact, many economies typically do have substantial labor inflows under *gastarbeiter* or other programs defined by immigration-quota policies, and, of course, equally there are substantial flows of capital among nation states.[17]

References

Bhagwati, Jagdish N. "Immiserizing Growth: A Geometrical Note." *Rev. Econ. Studies* 25 (June 1958): 201–5. (*a*)

———. "International Trade and Economic Expansion." *A.E.R.* 48 (December 1958): 941–53. (*b*)

———. "The Gains from Trade Once Again." *Oxford Econ. Papers* 20 (July 1968): 137–48.

———. "Shifting Comparative Advantage, Protectionist Demands, and Policy Response Options." Paper presented at a Ford-NBER conference on Import Competition and Adjustment: Theory and Policy, Cambridge, Mass., May 1980.

Bhagwati, Jagdish N., and Brecher, Richard A. "National Welfare in an Open Economy in the Presence of Foreign-owned Factors of Production." *J. Internat. Econ.* 10 (February 1980): 103–15.

Bhagwati, Jagdish N., and Kemp, Murray C. "Ranking of Tariffs under Monopoly Power in Trade." *Q.J.E.* 83 (May 1969): 330–35.

Bhagwati, Jagdish N.; Ramaswami, V. K.; and Srinivasan, T. N. "Domestic Distortions, Tariffs, and the Theory of Optimum Subsidy: Some Further Results." *J.P.E.* 77 (November/December 1969): 1005–10.

[17] Variations in protection may, in turn, lead to variations in the presence of foreign factors in the economy. Thus, for instance, Bhagwati (1980) has considered linkages between reduction in protection and reduction in the restrictiveness of immigration quotas. However, the analysis in the text has taken the endowment of *both* national and foreign-owned factors to be invariant to policy changes.

Bhagwati, Jagdish N., and Tironi, Ernesto. "Tariff Change, Foreign Capital and Immiserization: A Theoretical Analysis." *J. Development Econ.* 7 (March 1980): 71–83.

Johnson, Harry G. "The Cost of Protection and the Scientific Tariff." *J.P.E.* 68 (August 1960): 327–45.

Kemp, Murray C. *The Pure Theory of International Trade and Investment.* Englewood Cliffs, N.J.: Prentice-Hall, 1969.

Metzler, Lloyd A. "Tariffs, the Terms of Trade, and the Distribution of National Income." *J.P.E.* 57 (February 1949): 1–29.

Mundell, Robert A. "The Pure Theory of International Trade." *A.E.R.* 50 (March 1960): 67–110.

Rybczynski, T. M. "Factor Endowment and Relative Commodity Prices." *Economica* 22 (November 1955): 336–41.

Samuelson, Paul A. "The Transfer Problem and Transport Costs: The Terms of Trade When Impediments Are Absent." *Econ. J.* 62 (June 1952): 278–304.

———. "The Transfer Problem and Transport Costs, II: Analysis of Effects of Trade Impediments." *Econ J.* 64 (June 1954): 264–89.

Stolper, Wolfgang F., and Samuelson, Paul A. "Protection and Real Wages." *Rev. Econ. Studies* 9 (November 1941): 58–73.

Accounting for Price Changes: American Steel Rails, 1879–1910

Robert C. Allen

University of British Columbia

A framework is developed for decomposing product price changes into changes in input prices, technical efficiency, and deviations of price from unit cost. This framework facilitates the measurement of productivity growth in noncompetitive industries. The history of American steel rail prices between 1879 and 1910 is analyzed, and it is concluded (in contrast with much recent work) that productivity growth remained rapid until the twentieth century and that the steel industry was sufficiently collusive so that the rail producers received the benefits of that productivity growth as excess profits.

I

The 1890s mark a pronounced change in the position of the American steel industry in the world market. Before the 1890s American prices substantially exceeded British prices, and the American industry achieved a large size only because of high tariffs. During the 1890s American prices dropped to British levels or below, and America emerged as a major exporter of iron and steel. This change in price is apparent in the case of rails, which was one of the leading steel products of the time. Between 1881 and 1890 the average price of steel rails at Pennsylvania mills was $37.01 while the average British price was $23.62. During the period 1906–13 the American price had fallen to $28.00 while the British price had risen to $29.46 (Allen

I am extremely grateful to W. E. Diewert for his freely given assistance during the gestation period of this paper. I also thank Michael B. Percy, Peter Chinloy, C. F. J. Boonekamp, Frank Lewis, Peter Lindert, E. R. Berndt, and David Donaldson for their comments on earlier versions of the paper.

[*Journal of Political Economy*, 1981, vol. 89, no. 3]
© 1981 by The University of Chicago. 0022-3808/81/8903-0002$01.50

1979). To explain the emergence of the United States as a steel exporter, it is necessary to explain the fall in American steel prices.[1] It is the aim of this paper to account for the decline in the price of Bessemer steel rails.

There is no shortage of possible explanatory factors. During the late nineteenth century low-cost iron ore fields—particularly the Mesabi in Minnesota—were exploited, and improvements in lake shipping services were effected. Both developments substantially lowered the cost of iron ore in the Midwest and hence the cost of Bessemer pig iron, which was the principal input in steel rails. Moreover, many observers at the time believed the rate of productivity growth to have been very high in the American steel industry. This development would also be expected to have lowered the price of American steel. On the other hand, there were several attempts on the part of producers to collude and thus raise the price of steel rails. With the formation of U.S. Steel in 1901 the difficulty of enforcing collusive agreements was reduced. In the case of steel rails there is prima facie evidence that producers subsequently exercised monopoly power. Prior to 1901, the price of steel rails had changed frequently. In that year, however, the price of steel rails was set at $28 per ton, and it remained unchanged until the First World War. Such stability was unprecedented in the industry and indicates some power to control the price but does not in itself prove that price was set above cost (Temin 1964, pp. 192, 284–85).

The procedures that economic historians have used to analyze price changes have not been fully satisfactory. The technique most firmly rooted in theory is the application of an identity derived by Jorgenson (1966, pp. 2–4). The identity asserts that the rate of change of the price of the product equals an index of the rates of change of input prices minus a residual which is usually interpreted as the rate of technical progress. McCloskey (1973, p. 99) used a relative levels version of this identity to argue that—since the ratio of the price of rails to the price of pig iron (which was regarded as a proxy for the index of input prices) was trendless after the 1880s—there was no productivity growth after that date.[2] The difficulty with this procedure is that the identification of the residual with technical progress is correct only if the industry is in competitive long-run equilibrium. This assumption was barely mentioned in Jorgenson's discussion of the identity,[3] and, indeed, it is not immediately obvious from the

[1] It is also necessary to explain the rise in British prices. This aspect of the problem is considered in Allen (1979).

[2] Temin (1964, pp. 218–19) develops a similar argument.

[3] The discussion is confined to Jorgenson (1966, p. 4, n. 1).

mathematics that the assumption has been invoked. However, a scheme which decomposes price changes into efficiency and input price effects only if the industry is in competitive long-run equilibrium is not a satisfactory framework for analyzing American steel prices in the late nineteenth and early twentieth centuries.

Reflection suggests that changes in market power influence the difference between input and output prices. This realization has also been exploited by economic historians analyzing the evolution of steel prices. Thus McCloskey (1973, pp. 24–28) examined the ratio of rail to pig iron prices in order to investigate the effectiveness of various attempts at collusion by rail producers in Britain. Temin (1964, pp. 186–92) undertook similar investigations for the United States. Clearly, changes in monopoly power and technical progress cannot both be inferred from the ratio of output to input prices. At best, only their combined effect can be identified.

What is required to account for the fall in the price of American steel is an accounting scheme that identifies the separate effects of input price changes, efficiency changes, and changes in the deviation of price from unit cost. Such an accounting scheme is developed in Section II for a single product firm (an analogous scheme for a multiproduct firm is developed in an appendix).[4] In Section III the scheme is used to account for the fall in the price of steel rails between 1879 and 1910. The data discussed there indicate that declines in input prices and increases in efficiency both exerted powerful downward pressures. Indeed, the rate of technical progress was at its highest in the 1890s. The greater effectiveness of collusion after the U.S. Steel merger, however, meant that these cost declines were not entirely reflected in lower prices. Indeed, profit margins were raised enough between 1889 and 1902 so that rail producers realized most all of the gains of technical progress as higher profits. In Section IV these conclusions are compared with McCloskey's and Temin's findings. Jorgenson's price accounting identity is examined in detail in order to make explicit the necessity of assuming competitive long-run equilibrium if one tries to infer the rate of productivity change from differences in the rates of change of input and product prices. It is also shown that substantial errors result if this assumption is erroneously maintained.

[4] The author has extended the price accounting framework discussed here to the case of a multiple output firm. The multiple output case is developed in Allen (1978, pp. 7–11) and in an appendix to this paper which is available from the author upon request.

II

To develop the price accounting scheme for a single product firm, it is assumed that the firm minimizes costs subject to the following production function:

$$Q(t) = A(t) \cdot f[X(t)], \tag{1}$$

in which $Q(t)$ is output at time t, $X(t) = [X_1(t), X_2(t), \ldots, X_k(t)]$ is the vector of inputs used at time t, and $A(t)$ is an index of Hicks neutral technical progress. The assumption of Hicks neutrality can easily be relaxed (Diewert 1980, pp. 489–95). The function f is assumed to be linearly homogeneous. If f were known, the relative increase in efficiency between t_2 and t_1 could be computed from $Q(t_1)$, $Q(t_2)$, $X(t_1)$, and $X(t_2)$:

$$\frac{A(t_2)}{A(t_1)} = \frac{Q(t_2)/Q(t_1)}{f[X(t_2)]/f[X(t_2)]}. \tag{2}$$

The term $Q(t_2)/Q(t_1)$ is the relative increase in output that actually occurred. The term $f[X(t_2)]/f[X(t_1)] = A(t_1) \cdot f[X(t_2)]/A(t_1) \cdot f[X(t_1)]$ is the relative increase in output that would have occurred in the absence of technical change but given the change in input quantities that actually occurred. Thus $A(t_2)/A(t_1)$ equals the proportional increase in output that is not attributable to the increase in inputs.

If the firm minimizes cost, then there exists a total cost function C which is factorable into a unit cost function c^* multiplied by the production rate. Under the assumption that technical progress is Hicks neutral, the unit cost function can be written as a function of the input prices divided by $A(t)$:

$$C[t, W(t), Q(t)] = c^*[t, W(t)] \cdot Q(t) = c[W(t)] \cdot Q(t)/A(t). \tag{3}$$

Here $W(t) = [W_1(t), W_2(t), \ldots, W_k(t)]$ is the vector of k input prices. It is presumed that the supply to the firm of $X(t)$ is perfectly elastic at $W(t)$. The right-hand equality implies that the relative change in unit costs can be decomposed into an efficiency component and an input price component:

$$\frac{c^*[t_2, W(t_2)]}{c^*[t_1, W(t_1)]} = \frac{A(t_1)}{A(t_2)} \cdot \frac{c[W(t_2)]}{c[W(t_1)]}. \tag{4}$$

The term $A(t_2)/A(t_1)$ is relative efficiency in the two periods and $c[W(t_2)]/c[W(t_1)]$ is the relative increase in unit cost that would have obtained had there been no productivity growth but given the actual change in input prices. Equation (4) provides a second method for measuring productivity growth. If c were known, then $c[W(t_2)]/c[W(t_1)]$

could be computed by direct substitution. Dividing that result by the observed change in unit costs would yield $A(t_2)/A(t_1)$. From the perspective of equation (4), the increase in efficiency equals the relative decline in unit costs not attributable to declining input prices.

The price of the product will always be related to unit costs by the following identity:

$$P(t) = M(t) \cdot c^*[t, W(t)], \tag{5}$$

where $M(t)$ is the implicitly defined markup of price over cost. For a competitive industry in long-run equilibrium, $M(t) = 1$. One would expect to observe $M(t)$ usually not equal to one. The relative change in price between t_1 and t_2 can be written as

$$\frac{P(t_2)}{P(t_1)} = \frac{M(t_2)}{M(t_1)} \cdot \frac{c^*[t_2, W(t_2)]}{c^*[t_1, W(t_1)]}. \tag{6}$$

Equation (6) could be used to decompose a change in price into a cost change and a markup change. Equations (2) and (4) provide a basis for decomposing the cost change into input price and efficiency components.

If c and f were known, equations (2), (4), and (6) could be used to decompose price changes into markup, efficiency, and input price changes. In an empirical study, of course, the functional forms of c and f would have to be specified and the parameters of the function estimated. Prodigious task! Recent work in index number theory (e.g., Diewert 1976, 1980) has shown how such elaborate undertakings can be circumvented. Corresponding to each of the functional forms that would likely be used to specify c or f are one or more index number formulae. The indexed input quantities (or input prices) would indicate exactly $f[X(t_2)]/f[X(t_1)]$ or $c[W(t_2)]/c[W(t_1)]$ on the assumption that f or c had the functional form that corresponded to the index number formula used. For example, $f[X(t_2)]/f[X(t_1)]$ would equal a Laspeyres or Paasche index of the input quantities if f were Leontief. If f were Cobb-Douglas, $f[X(t_2)]/f[X(t_1)]$ would equal a geometric index of input quantities. If f were a quadratic function, $f[X(t_2)]/f[X(t_1)]$ would equal a Fisher ideal index, and if f were translog, then $f[X(t_2)]/f[X(t_1)]$ would equal a Törnqvist index of the input quantities.[5] Analogous results

[5] These equivalences are discussed in Diewert (1980, pp. 446–52). The linearly homogeneous translog function, which is used extensively in this paper, is

$$\ln f[X(t)] = \alpha_0 + \sum_{i=1}^{k} \beta_i \ln X_i(t) + 1/2 \sum_{m=1}^{k} \sum_{n=1}^{k} \beta_{mn} \ln X_m(t) \ln X_n(t),$$

subject to the restrictions. $\Sigma_{i=1}^{k}\beta_i = 1$, $\beta_{mn} = \beta_{nm}$, and $\Sigma_{i=1}^{k}\beta_{mn} = 0$ for $m = 1, 2, \ldots, k$. Diewert (1976) has shown that

hold for c and input price indices. It should be noted that in the index number approach to the measurement of $f[X(t_2)]/f[X(t_1)]$ or $c[W(t_2)]/c[W(t_1)]$ it is not necessary to know the values of the parameters of f or c. Instead, the index number formulae require the values of the factor shares, estimates of which are frequently easier to obtain than estimates of the parameters of f or c.

In an empirical study, one should choose the formula for indexing input prices or quantities that corresponds to the true structure of c and f. In the absence of any prior information on the true structure, the choice of an index number formula must be made on the basis of which possible specification of f or c would best approximate the true structure, whatever it is. Since Leontief and Cobb-Douglas functions can be regarded as first-order approximations to f or c, while quadratic and translog functions can be regarded as second-order approximations, one would expect the quadratic and translog functions to fit the data more closely. Hence, there are grounds for preferring the Törnqvist and the Fisher ideal indices to the Laspeyres, Paasche, or geometric. Since the Fisher ideal and Törnqvist indices are exact for functions that can be regarded as second-order approximations to any linearly homogeneous production, cost, or expenditure function, Diewert (1976) has termed the Törnqvist and Fisher ideal indices "superlative." In the remainder of this paper, translog functions and Törnqvist indices will be used.

There are two ways to proceed, depending on whether one measures $A(t_2)/A(t_1)$ with equation (2) or (4). If one uses equation (2), one presumes f is translog, so $f[X(t_2)]/f[X(t_1)]$ equals a Törnqvist index of inputs. Hence,

$$\frac{A(t_2)}{A(t_1)} = \frac{Q(t_2)/Q(t_1)}{\prod_{i=1}^{k}\left[\frac{X_i(t_2)}{X_i(t_1)}\right]^{\bar{s}_i}} = \prod_{i=1}^{k}\left[\frac{Q(t_2)/X_i(t_2)}{Q(t_1)/X_i(t_1)}\right]^{\bar{s}_i}. \tag{7}$$

Consistency requires that $c[W(t_2)]/c[W(t_1)]$ be measured implicitly by substituting $A(t_2)/A(t_1)$ from equation (7) into equation (5):

$$\bar{c}[W(t_2),W(t_1)] = \frac{c^*[t_2,W(t_2)]}{c^*[t_1,W(t_1)]} \cdot \prod_{i=1}^{k}\left[\frac{Q(t_2)/X_i(t_2)}{Q(t_1)/X_i(t_1)}\right]^{\bar{s}_i}. \tag{8}$$

$$\frac{f[X(t_2)]}{f[X(t_1)]} = \prod_{i=1}^{k}\left[\frac{X_i(t_2)}{X_i(t_1)}\right]^{1/2[s_i(t_2)+s_i(t_1)]},$$

which is a Törnqvist index of $X(t)$. Here, $s_i(t)$ is the share in costs of $X_i(t)$.

Rearranging equation (8) indicates that the change in unit costs can be expressed as

$$\frac{c^*[t_2,W(t_2)]}{c^*[t_1,W(t_1)]} = \bar{c}[W(t_2),W(t_1)]/\prod_{i=1}^{k}\left[\frac{Q(t_2)/X_i(t_2)}{Q(t_1)/X_i(t_1)}\right]^{\bar{s}_i}. \tag{9}$$

Substituting equations (7) and (9) into equation (6) provides one scheme for decomposing product price changes into input, efficiency, and markup components:

$$\frac{P(t_2)}{P(t_1)} = \frac{M(t_2)}{M(t_1)} \cdot \frac{A(t_1)}{A(t_2)} \cdot \bar{c}[W(t_2),W(t_1)]. \tag{10}$$

If one uses equation (4) to measure productivity growth, a slightly different accounting scheme emerges. In this case, one presumes c is translog, so $c[W(t_2)]/c[W(t_1)]$ equals a Törnqvist index of input prices:

$$\frac{c[W(t_2)]}{c[W(t_1)]} = \prod_{i=1}^{k}\left[\frac{W_i(t_2)}{W_i(t_1)}\right]^{\bar{s}_i}. \tag{11}$$

Equation (4) can then be used to implicitly define an efficiency index:

$$\tilde{A}(t_2,t_1) = \prod_{i=1}^{k}\left[\frac{W_i(t_2)}{W_i(t_1)}\right]^{\bar{s}_i} \Big/ \frac{c^*[t_2,W(t_2)]}{c^*[t_1,W(t_1)]}. \tag{12}$$

Analogously to equation (9), one can decompose unit cost changes as:

$$\frac{c^*[t_2,W(t_2)]}{c^*[t_1,W(t_1)]} = \prod_{i=1}^{k}\left[\frac{W_i(t_2)}{W_i(t_1)}\right]^{\bar{s}_i}/\tilde{A}(t_2,t_1). \tag{13}$$

Substituting equation (13) into equation (6) provides the second price accounting scheme:

$$\frac{P(t_2)}{P(t_1)} = \frac{M(t_2)}{M(t_1)}\prod_{i=1}^{k}\left[\frac{W_i(t_2)}{W_i(t_1)}\right]^{\bar{s}_i}/\tilde{A}(t_2,t_1). \tag{14}$$

Equations (10) and (14) are each internally consistent but differ in the presumptions they make about the way the data were generated. In equation (10) the presumption is that the data were generated by minimizing costs subject to a translog production function. In equation (14) the presumption is that costs were minimized subject to the production function which generates a translog cost function. Since a translog production function does not yield a translog cost function (and conversely), different accounting schemes follow from the different assumptions. However, since translog production and cost functions both provide second-order approximations to the technology, one would expect the two accounting schemes to give similar results. That, indeed, is true in the case of steel rails.

III

Table 1 presents details of the costs of producing steel rails in the United States in 1879, 1889, 1902, and 1910. The figures for the latter two years are derived from the U.S. Commissioner of Corporations (1913) investigation into costs, prices, and profits in the steel industry. The commissioner obtained and published details of the cost of producing most steel products made in the United States between 1902 and 1910. The figures for 1902 are averages for firms producing virtually all of the heavy steel rails made in America in that year. The 1910 figures are averages for all U.S. Steel's Bessemer rail mills. These mills produced 60.2 percent of American production in that year. The cost figures for 1901 and 1903–9 that the bureau reports are quite similar to the 1902 and 1910 figures, so conclusions founded on those years are indicative of conditions throughout the first decade following the U.S. Steel merger.[6]

The cost figures for 1879 and 1889 are less certain than those for 1902 and 1910. The 1889 figures are based on the U.S. Commissioner of Labor's (1891) survey of steelmaking costs in the United States, Great Britain, and the European continent. Although the cost estimates were based on the cost accounts of steel firms, the coverage was not as broad as the U.S. Bureau of Corporations survey. Moreover, the U.S. Commissioner of Labor (1891, p. 163) *Report* expressed concern that some of the leading American rail mills had refused to supply cost information. The 1879 figures are also from an international comparison of production costs. This survey was conducted by Alexander L. Holley, who was the leading American steelworks engineer. The survey was published privately by the Bessemer Association, which owned the American rights to the Bessemer patents, for the use of its licensees. Holley visited several European plants in 1879 and reported their costs. For comparison, he provided "representative costs" in the United States. It is the latter costs which are used in table 1. Given Holley's knowledge, the audience he was writing for, and the confidential nature of the communication, one presumes the American cost figures are indeed representative.[7]

[6] See U.S. Commissioner of Corporations (1913, pp. 209–10, 461) for 1910 rail production figures and pp. 143–49, 209–16, 346–68 for information on costs in the other years.

[7] It should be noted that none of these sources reports the quantity and price of all inputs consumed. That practice was followed in the case of metallic inputs, but in the cases of fuel, labor, miscellaneous materials, and capital, only the expense per ton was recorded. It was necessary to estimate input prices and implicitly input quantities from other sources. Those sources are discussed in the notes to table 1. It must also be stressed that all calculations reported in this paper assume that the costs in table 1 are minimum long-run average total costs.

TABLE 1

RAIL COSTS, PRICES, AND MARKUPS

INPUT	1879			1889			1902			1910		
	W_i	X_i/Q	W_iX_i/Q	W_i	X_i/Q	W_iX_i/Q	W_i	X_i/Q	W_iX_i/Q	W_i	X_i/Q	W_iX_i/Q
Metal	20.60	1.29	26 47	15.21	1.24	18.86	14 78	1.09	16.18	15.36	1.09	16.67
Fuel	1.06	2.08	2.20	1.04	2.04	2.13	1.68	.54	.91	1 55	.50	.77
Labor	.432	8 472	3 66	.577	5.806	3.35	.609	3.284	2.00	.697	2.396	1.67
Materials	.90	1.90	1.77	.81	1.85	1.50	.859	1.69	1.45	1.027	1.09	1.12
Capital	7.1848	.1851	1.33	5.8252	.2263	1.32	5.7942	.2675	1.55	6.2888	.2067	1.30
Average total cost	..	..	35.37	...	...	27 16	..	...	22.09	. .	...	21.53
Price of rails	...	...	40.18	...	.	29.25	..	...	28.00	..	...	28.00
M	...	...	1.14	.	...	1.08	..	..	1.27	...	...	1.30

SOURCES.—Except as noted subsequently the price (dollars per long ton) and quantity (long tons) consumed of metal (which includes manganese as well as pig iron and any scrap) and the expenses of the various inputs per ton of rails were computed from Holley (1881), U S Commissioner of Labor (1891, p 173), and U S Commissioner of Corporations (1913, pp. 143–49, 209–16, 426–29, 460–68) In all cases it was necessary to combine separately reported costs for the production of ingots, the rolling of billets, and the rolling of rails It was also necessary to estimate the capital cost component of the pig iron price for 1889 This cost was estimated as 10% of the capital invested in American blast furnaces in 1889 divided by the production of pig iron These figures were taken from U S Census Office (1902, p 28) Expenses for other inputs were broken into prices and quantities as follows. *Fuel* The price of fuel (dollars per long ton) was taken to be the price of bituminous coal purchased by steelworks and rolling mills For 1879 the price of coal is the price paid by Pennsylvania steelworks as given in U.S Census Office (1883, p. 757) deflated by the ratio of the 1879 to 1880 value of the fuel and lighting component of the Warren and Pearson price index given in U S. Bureau of the Census (1975, 1.201) The low price of metallic inputs reported by Holley and the dates at which he visited European works suggest he was reporting early 1879 American values. In mid-1879 American steel prices and steel input prices rose, and this rise is reflected in the 1879 census figures which were collected for the year June 1 1879–May 31 1880 Hence, the fuel price in the census was deflated The U.S Commissioner of Labor (1891) reported the quantity of coal consumed per ton of rails. The price of coal was taken to equal coal expense per ton of rails divided by fuel consumption. For 1902 and 1910 the price of fuel was taken to equal the price of coal consumed by steelworks and rolling mills in 1899 (given in U.S. Census Office 1902, p. 57), inflated in proportion to the rise in the fuel and lighting component of the BLS wholesale price index given in U S Bureau of the Census (1975, 1 200) For 1902 and 1910 fuel expenses in table 1 include the cost of steam The quantity of fuel (in long tons) consumed equals fuel expense divided by the price of fuel *Labor* The price of labor (thousands of dollars per man-year) is the average earnings of workers employed in steelworks and rolling mills For 1879 and 1889 the average is computed from U.S Bureau of the Census (1913, p. 228) For 1902 and 1910 the average is taken from Douglas (1930, p. 271) The quantity of labor equals labor expense divided by the price of labor *Materials* The price of materials is taken to be the Warren and Pearson price index extended past 1890 by the BLS wholesale price index (U S Bureau of the Census 1975, 1 200–201). These have been divided by 100 The quantity of materials equals material expense divided by the price of materials. *Capital* The price of capital equals the deflator for capital in the iron and steel industry developed by Creamer, Dobrovolsky, and Borenstein (1960, p 260) multiplied by an interest rate plus a depreciation rate The deflators given for 1880, 1890, 1900, and 1909 were applied to the years 1879, 1889, 1902, and 1910 respectively The depreciation rate was presumed to be .059 on the basis of Creamer et al (1960, p 223) The interest rate was taken to be the Cowles Commission yield on industrial common stock as given by U S Bureau of the Census (1975, 2 1003) The quantity of capital equals capital expense divided by the price of capital The U S. Commissioner of Labor (1891) did not report any capital cost information so the quantity of capital in 1889 was estimated as the average of the quantities of capital in 1879 and 1902. Capital expense in 1889 then equals that quantity multiplied by the 1889 price of capital *Price of rails* The price of rails was taken from U S. Commissioner of Labor (1891, p. 179) and Temin (1964, pp. 284–85). For reasons given in the discussion of fuel prices, an early 1879 price of pig iron ($41 per ton) was chosen. The $40.18 shown in the table equals $41 less the royalty of $ 82 per ton of rails ($.75 per ton of ingots). The royalty lapsed by 1889 and subtracting it from the price in 1879 seemed the easiest way to incorporate it.

NOTE.—W_i = the price of input i X_i/Q = the consumption of input i per ton of rails, and W_iX_i/Q = the expense per ton of rails of input i Average total cost equals the sum of W_iX_i/Q for each year M = the price of rails divided by the average cost

Between the 1880s and early 1900s the American steel industry became internationally competitive because the price of steel in the United States fell sharply. Table 1 and tables 2 and 3, which report the index numbers discussed in the previous section, make clear the decline in price was associated with a sharp decline in steelmaking costs. According to table 2, steelmaking costs in 1889 equaled 77 percent of costs in 1879. In 1902 costs were 81 percent of their 1889 level. The decline in costs subsequently slowed, for costs in 1910 equalled 97 percent of their 1902 level.

Equations (9) and (13) provide schemes for decomposing the cost changes into efficiency and input price components. The results from using both equations are exhibited in table 2, and, as the reader can verify, the results are nearly identical. Between 1879 and 1889 falling input prices were the main cause of falling costs. Thereafter input prices rose slowly. Between 1889 and 1902 costs continued to fall because productivity growth was rapid. After 1902, the rate of productivity growth slackened and input prices rose, so the decrease in costs between 1902 and 1910 was meager.

Equation (7) indicates that productivity increased at the rate of 0.6 percent per year between 1879 and 1889. Between 1889 and 1902 the rate increased to 1.6 percent and then declined to 0.8 percent in the period 1902–10. (Eq. [12] gives essentially the same results.) Throughout the period 1879–1910, the average product of labor continuously rose, and the average product of capital generally fell. Since the ratio of the wage rate to the rental price of capital consistently increased, the change in the average products of labor and capital might have been due to factor substitution. Moreover, recent econometric work suggests that technical change was labor augmenting in this period (Cain and Paterson 1979). Technical change and factor substitution reinforced each other in determining the labor-capital mix.

Productivity growth was highest between 1889 and 1902. What distinguished those years from the others was the marked rise in the average products of fuel and metallic inputs. Fuel consumption declined as the mixer (invented in 1889) was adopted by American steelworks. In earlier years, it had been impossible to use the molten pig iron from blast furnaces directly in Bessemer converters. The pig iron was cast and then remelted. By 1902 most Bessemer steel was made from molten iron conveyed directly from the blast furnace. The saving in fuel was impressive. The productivity of metallic inputs also increased due to developments in other areas of the steel mill, in this case due to the growth of open hearth steelmaking capacity. Bessemer shops and rolling mills generated a large volume of scrap in the form of poor rolls and sheared ends. These could not be recycled in

TABLE 2

DECOMPOSING UNIT COST CHANGES

	By Eq. (9)					By Eq. (13)				
Period	$\frac{c^*[t_2,W(t_2)]}{c^*[t_1,W(t_1)]}$	=	$\frac{A(t_1)}{A(t_2)}$	·	$\bar{c}[W(t_2),W(t_1)]$	$\frac{c^*[t_2,W(t_2)]}{c^*[t_1,W(t_1)]}$	=	$\frac{1}{\bar{A}(t_2,t_1)}$	·	$\frac{c[W(t_2)]}{c[W(t_1)]}$
1879–89	.7679	=	(.9367)	·	(.8198)	.7679	=	(.9397)	·	(.8172)
1889–1902	.8133	=	(.7962)	·	(1.0215)	.8133	=	(.7994)	·	(1.0174)
1902–10	.9746	=	(.9305)	·	(1.0475)	.9746	=	(.9242)	·	(1.0545)

SOURCE.—Table 1. $c^*[t_2,W(t_2)]/c^*[t_1,W(t_1)]$ is observed unit cost (from table 1) in the terminal year of the period divided by observed unit cost in the initial year. $A(t_1)/A(t_2)$ was computed according to eq. (7), $\bar{c}[W(t_2),W(t_1)]$ according to eq. (8), $c[W(t_2)]/c[W(t_1)]$ according to eq. (11), and $\bar{A}(t_2,t_1)$ according to eq. (12).

TABLE 3

DECOMPOSING PRICE CHANGES

	By Eq. (10)						By Eq. (4)					
Period	$\frac{P(t_2)}{P(t_1)}$	=	$\frac{M(t_2)}{M(t_1)}$	·	$\frac{A(t_1)}{A(t_2)}$	$\bar{c}[W(t_2),W(t_1)]$	$\frac{P(t_2)}{P(t_1)}$	=	$\frac{M(t_2)}{M(t_1)}$	·	$\frac{1}{\bar{A}(t_2,t_1)}$	· $\frac{c[W(t_2)]}{c[W(t_1)]}$
1879–89	.7280	=	(.9480)	·	(.9367)	· (.8198)	.7280	=	(.9480)	·	(.9397)	· (.8172)
1889–1902	.9573	=	(1.1771)	·	(.7962)	· (1.0215)	.9573	=	(1.1771)	·	(.7994)	· (1.0174)
1902–10	1.000	=	(1.0261)	·	(.9305)	· (1.0475)	1.000	=	(1.0261)	·	(.9242)	· (1.0545)

SOURCE.—Table 1. $P(t_2)/P(t_1)$ is the ratio of the price in the terminal year of the period to the price in the initial year. See table 2 for the definitions of other variables.

Bessemer converters but could be remelted in open hearth furnaces. In 1889 only 11 percent of American steel was open hearth, but by 1902 38 percent was open hearth (Temin 1964, pp. 272–73). The development of this process allowed the recycling of scrap. Tables 1 and 2 follow the nineteenth-century accounting tradition of charging Bessemer rail mills for their net consumption of metallic inputs. The recycling of Bessemer scrap in open hearth plants thus shows up as a rise in the productivity of metallic inputs in Bessemer rail production.

Table 1 indicates that the fall in production costs was not entirely passed on to consumers in lower prices. An apparent consequence of the U.S. Steel merger was to raise markups to the 30 percent range for at least the first decade after the merger. In 1879 and 1889 markups had been much lower.

Table 3 uses equations (10) and (14) to decompose price changes into markup, efficiency, and input price components. The price of rails in 1889 equalled 72.8 percent of the price in 1879. Some of the decline in price was due to a decline in the markup; however, most of the decline can be ascribed to falling costs. As indicated previously, declining input prices were the predominant cause of declining costs. The fall in input prices was also decisive for the decline in price.

Between 1889 and 1902, costs declined 19 percent but rail prices declined only 4 percent. The discrepancy arose because markups rose 18 percent. This result indicates a major consequence of the U.S. Steel merger. Steel prices in the early twentieth century were set at levels prevailing about 1890, and the steel producers absorbed all of the decline in costs (due to technical progress in the case of rails) as excess profits.

Table 3 indicates that there was little change in price, costs, or the determinants of costs between 1902 and 1910. The steel industry was sufficiently collusive so that a markup of 30 percent over unit costs could be maintained for a decade.

IV

These conclusions regarding productivity growth and the market power exercised after the U.S. Steel merger are at variance with recent conclusions advanced by McCloskey (1973, p. 99) and Temin (1964, pp. 186–93, 218–20). Both noted that the ratio of the price of steel rails to the price of pig iron was trendless from the 1880s to the First World War. (This constancy is illustrated by the data in table 1.) Both inferred from these price data that productivity growth ceased in the 1880s, and Temin also inferred that market control after the U.S. Steel merger was not effective in raising the price of rails.

To choose between McCloskey's and Temin's conclusions and those

advanced here, it is necessary to scrutinize the theoretical underpinnings of the calculations employed. McCloskey's procedure consists of the (approximate) application of an identity deduced by Jorgenson (1966, pp. 2–4). Although the assumption is not obviously apparent in the derivation, it is the case that this procedure only correctly measures productivity growth (and, hence, accounts for price changes) for industries in competitive long-run equilibrium. This will be shown subsequently. It is worth considering the matter thoroughly, both to resolve the issue at hand and because Jorgenson's identity has been widely used by economic historians to measure productivity growth and account for price changes.[8]

Jorgenson begins with the accounting identity that a firm's revenues necessarily equal the total income earned by the firm's inputs. In the case of a single product firm,

$$PQ = \sum_{i=1}^{k} W_i^* X_i, \tag{15}$$

where W_i^* are the accounting valuations of the inputs X_i. Totally differentiating equation (15) with respect to time, dividing the result by equation (15), and regrouping terms gives:

$$\frac{\dot{Q}}{Q} - \sum_{i=1}^{k} s_i^* \frac{\dot{X}_i}{X_i} = \sum_{i=1}^{k} s_i^* \frac{\dot{W}_i^*}{W_i} - \frac{\dot{P}}{P}. \tag{16}$$

Here the shares are $s_i^* = W_i^* X_i / \Sigma W_i^* X_i$.

Jorgenson defined productivity to equal the ratio of output to total input: $E \equiv Q/X$. Totally differentiating this expression with respect to time implies that the rate of change of efficiency equals the rate of change of output less the rate of change of total input: $\dot{E}/E = \dot{Q}/Q - \dot{X}/X$. Defining $\dot{X}/X$ to equal $\Sigma s_i^*(\dot{X}_i/X_i)$ implies

$$\frac{\dot{E}}{E} = \frac{\dot{Q}}{Q} - \sum_{i=1}^{k} s_i^* \frac{\dot{X}_i}{X_i} = \sum_{i=1}^{k} s_i^* \frac{\dot{W}_i}{W_i} - \frac{\dot{P}_i}{P_i}. \tag{17}$$

Jorgenson contended, moreover, that Solow (1957) deduced the left-hand equality from a production model.

If correct, equation (17) has an important implication for produc-

[8] North (1968, p. 954), Shepherd and Walton (1972, pp. 60–63), Harley (1973, p. 376), and McCloskey (1973, p. 86) cite Jorgenson (1966). Sandberg (1974, p. 100) cites the thesis version of McCloskey (1973) for a proof of the equivalence of the G. T. Jones index of real cost and the right-hand side of eq. (16) of this paper and goes on to assert the equivalence of that index and Solow's (1957) index. Most of these economic historians recognize that the identity is not a sound basis for measuring productivity change when long-run competitive equilibrium does not prevail, and a number of ad hoc adjustments are developed to deal with the problem. Lindert and Trace (1971) do develop productivity indices that exclude excess profits.

tivity measurement, namely, that productivity growth can be inferred from the growth of input prices less the growth in output price *without any explicit account being taken of markups.* This result justified McCloskey's procedure and appears inconsistent with equations (10) and (14) in this paper.

Equation (17) is an identity and hence always true. The difficulty with using it as the basis of productivity measurement is that $\dot{E}/E$ equals the rate of productivity growth in the Solow (1957) sense only if the firm (or industry) whose productivity growth is being measured is in long-run, competitive equilibrium. This condition was not made explicit in Jorgenson's derivation. To see the importance of the competitive equilibrium assumption, recall Solow's derivation. Solow (1957) assumes a production function of the form of equation (1) and identifies the rate of productivity growth with $\dot{A}/A$. Totally differentiating equation (1) with respect to time, dividing the result by equation (1), and rearranging allows $\dot{A}/A$ to be expressed as

$$\frac{\dot{A}}{A} = \frac{\dot{Q}}{Q} - \sum_{i=1}^{k} \frac{(\partial f/\partial X_i)X_i}{f(X)} \cdot \frac{\dot{X}_i}{X_i}, \tag{18}$$

where the time symbol t is suppressed for simplicity. If the firm faces exogenous input prices and minimizes costs, then the necessary conditions for cost minimization can be rewritten as $W_i/\Sigma W_i X_i = (\partial f/\partial X_i)/f(X)$. In that case equation (18) can be rewritten as

$$\frac{\dot{A}}{A} = \frac{\dot{Q}}{Q} - \sum_{i=1}^{k} s_i \frac{\dot{X}_i}{X_i}, \tag{19}$$

where $s_i = W_i X_i/\Sigma W_i X_i$.

Equation (19) and the left-hand equality of equation (17) are almost identical. They would be identical if $s_i^* = s_i$. Since equation (17) is an accounting identity, one must make some assumptions about the institutional and accounting framework in order to determine if $s_i^* = s_i$. Reasonable, simple assumptions are that inputs are available to the firm in infinitely elastic supply at prices W_i, any difference between revenues and the opportunity cost of inputs accrues as excess profit (or loss) to capital, and all inputs besides capital are valued in the firm's accounts at their supply prices W_i. Given these presumptions $W_i^* = W_i$ for all inputs besides capital. For capital (which will be indexed as the kth input), $W_k^* \gtreqless W_k$ if revenues exceed, equal, or fall short of the opportunity cost of the inputs.

If the firm is in competitive, long-run equilibrium, revenues equal the opportunity cost of all the inputs and $W_k^* = W_k$. In this case $s_i^* = s_i$, and equation (17) correctly measures the rate of productivity growth $\dot{A}/A$. Moreover, in this case, productivity growth can be inferred from

input and product price movements without taking markups into account. This result is consistent with the framework of Section II of this paper, for, if revenues always equal opportunity costs, $M(t_1) = M(t_2) = 1$ and the markup terms disappear from equation (14). Efficiency change can then be inferred from changes in product and input prices.

Suppose, on the other hand, that the markup exceeds one. In that case revenues exceed opportunity costs, $W_k^* > W_k$, and $\Sigma W_i^* X_i > \Sigma W_i X_i$. Hence, $s_k^* > s_k$ and $s_i^* < s_i$ for all inputs besides capital. Unless all inputs (including capital) grow at the same rate $\Sigma s_i^*(\dot{X}_i/X_i) \neq \Sigma s_i(\dot{X}_i/X_i)$ and the rate of productivity growth cannot be inferred from equation (17). The same conclusion follows if markups are less than one. Therefore, if markups do not equal one, equation (17) cannot be used to measure the rate of shift of the production function. In particular, one cannot infer the rate of productivity growth from differences in the rates of growth of input and product prices.

Substantial errors result if equation (17) is used when markups do not equal one. The errors are apparent if equation (17) is applied to the data in table 1. That application requires that some discrete approximation be adopted. The most useful approximation is to approximate the time-varying shares s_i^* by the average of the shares prevailing at times t_1 and t_2, $\bar{s}_i^* = 1/2\,[s_i^*(t_1) + s_i^*(t_2)]$, and integrate the left-hand equality in equation (17) between t_1 and t_2. Exponentiating the integral yields:

$$\frac{E(t_2)}{E(t_1)} = \prod_{i=1}^{k} \left[\frac{Q(t_2)/X_i(t_2)}{Q(t_1)/X_i(t_1)} \right]^{\bar{s}_i^*}. \tag{20}$$

If one treats equation (19) analogously, the result is equation (7). Again the only difference between the putative productivity measures is in the shares. Numerically, however, the differences are not inconsequential. For the periods 1879–89, 1889–1902, and 1902–10 $A(t_2)/A(t_1)$ equals 1.0676, 1.2560, and 1.0747, respectively. For the same periods $E(t_2)/E(t_1)$ equals 1.0408, 1.1904, and 1.1199. The term $E(t_2)/E(t_1)$ understates the average annual rate of productivity growth by 40 percent in 1879–89 and 24 percent in 1889–1902 and overstates the rate by 59 percent in 1902–10. Assuming long-run, competitive equilibrium when it does not obtain leads to serious errors in productivity measurement.

V

In this paper it has been shown that American rail prices fell in the late nineteenth century because production costs declined sharply. Two unexpected findings include the discovery that the rate of pro-

ductivity growth remained high into the twentieth century and the discovery that rail producers managed to maintain steel prices at a level 30 percent above average total cost for a decade after the U.S. Steel merger. Temin's (1964, pp. 189–93) recent work on the merger concluded that the markets for rolled products remained competitive after the merger. That conclusion is certainly not true for rails. Similar evidence for other products suggests that markups were generally high.[9] A reassessment of the significance of the U.S. Steel merger in this regard seems in order.

In order to establish these conclusions, it was necessary to develop a more elaborate price accounting scheme than those previously available. As part of that effort, considerable attention was devoted to the circumstances under which one could infer the rate of productivity change from the difference in the rates of input price and product price changes. A correct inference can be made only if the industry is in competitive long-run equilibrium. If that condition does not obtain, serious errors are likely to result. These difficulties can be avoided if productivity growth is measured with equation (7) or (12).

References

Allen, Robert C. "Accounting for Price Changes." Discussion Paper no. 78-14, Univ. British Columbia, Dept. Econ., 1978.

———. "International Competition in Iron and Steel, 1850–1913." *J. Econ. Hist.* 39 (December 1979): 911–37.

Cain, L. P., and Paterson, D. G. "Factor Biases and Technical Change in Manufacturing: The American System, 1850." Mimeographed, Univ. British Columbia, Dept. Econ., 1979.

Creamer, Daniel; Dobrovolsky, Sergei P.; and Borenstein, Israel. *Capital in Manufacturing and Mining: Its Formation and Financing.* Princeton, N.J.: Princeton Univ. Press, 1960.

Diewert, W. E. "Exact and Superlative Index Numbers." *J. Econometrics* 4 (May 1976): 115–45.

———. "Aggregation Problems in the Measurement of Capital." In *The Measurement of Capital*, edited by Dan Usher. Chicago: Univ. Chicago Press (for Nat. Bur. Econ. Res.), 1980.

Douglas, Paul H. *Real Wages in the United States, 1890–1926.* Boston: Houghton Mifflin, 1930.

Harley, C. K. "On the Persistence of Old Techniques: The Case of North American Wooden Shipbuilding." *J. Econ. Hist.* 33 (June 1973): 372–98.

Holley, Alexander L. *Holley's Reports to the Bessemer Steel Co. Limited: Bessemer and Rolling Mill Practice and Costs at the West Cumberland Iron & Steel Works and at Brown, Bailey & Dixon's Works.* New York: Russell Bros., 1881.

[9] U.S. Commissioner of Corporations (1913, pp. 520–31) shows substantial profits, but the figures as presented are not directly comparable with those in this paper and do not throw light on excess profits in the rolled product stage of the industry. Preliminary reworking of these figures using the methods developed here shows high markups for most rolled products.

Jorgenson, Dale W. "The Embodiment Hypothesis." *J.P.E.* 74 (February 1966): 1–17.

Lindert, Peter H., and Trace, Keith. "Yardsticks for Victorian Entrepreneurs." In *Essays on a Mature Economy: Britain after 1840*, edited by Donald N. McCloskey. Princeton, N.J.: Princeton Univ. Press, 1971.

McCloskey, Donald N. *Economic Maturity and Entrepreneurial Decline: British Iron and Steel, 1870–1913*. Cambridge, Mass.: Harvard Univ. Press, 1973.

North, Douglass C. "Sources of Productivity Change in Ocean Shipping, 1600–1850." *J.P.E.* 76 (September/October 1968): 953–70.

Sandberg, Lars G. *Lancashire in Decline: A Study in Entrepreneurship, Technology, and International Trade*. Columbus: Ohio State Univ. Press, 1974.

Shepherd, James F., and Walton, Gary M. *Shipping, Maritime Trade, and the Economic Development of Colonial North America*. Cambridge: Cambridge Univ. Press, 1972.

Solow, Robert M. "Technical Change and the Aggregate Production Function." *Rev. Econ. and Statis.* 39 (August 1957): 312–20.

Temin, Peter. *Iron and Steel in Nineteenth-Century America: An Economic Inquiry*. Cambridge, Mass.: MIT Press, 1964.

U.S. Bureau of the Census. *Thirteenth Census of the United States*. Vol. 10. *Manufactures. Reports for Principal Industries*. Washington: Government Printing Office, 1913.

———. *Historical Statistics of the United States*. Washington: Government Printing Office, 1975.

U.S. Census Office. *Report on the Manufactures of the United States at the Tenth Census*. Washington: Government Printing Office, 1883.

———. *Twelfth Census of the United States*. Vol. 10. *Manufactures: Part IV, Special Reports on Selected Industries*. Washington: Government Printing Office, 1902.

U.S. Commissioner of Corporations. *Report on the Steel Industry*. Pt. 3, *Cost of Production*. Washington: Government Printing Office, 1913.

U.S. Commissioner of Labor. *Sixth Annual Report: Cost of Production: Iron, Steel, Coal, etc.* Rev. ed. Washington: Government Printing Office, 1891.

Some Evidence on Cross-Sector Effects of the Minimum Wage

George E. Tauchen

Duke University

This paper tests Mincer's minimum-wage model by estimating reduced-form wage and employment equations for both the covered and uncovered sectors in nine regions of the United States. As theory predicts, in regions with comparatively small covered-sector demand elasticities, the northern and midwestern regions, the uncovered-sector wage increases after a minimum-wage hike; and in regions with comparatively large demand elasticities, the southern and western regions, the uncovered-sector wage decreases. Because of data limitations the uncovered-sector employment effect could not be estimated sharply, and so its relationship to the covered-sector demand elasticity is weak.

I. Introduction

Economic theory predicts that a minimum-wage hike decreases employment of low-skill workers in the covered sector. Theory also suggests that a minimum-wage hike affects the equilibrium wage and the level of employment in the uncovered sector though, as pointed out by Mincer (1976), the directions of the uncovered-sector effects are ambiguous. On the one hand, if the absolute labor demand elasticity is relatively small so that firms eliminate only a few covered jobs after the wage hike, then workers find the minimum wage attractive enough to justify the risk of unemployment while searching for jobs in the covered sector. In this case labor leaves the uncovered sector, and the wage thereby increases. On the other hand, if the

I am grateful to Christopher A. Sims, Helen Tauchen, and an anonymous referee for many helpful comments and suggestions.

[*Journal of Political Economy*, 1981, vol. 89, no. 3]
© 1981 by The University of Chicago 0022-3808/81/8903-0008$01.50

demand elasticity is relatively large so that the return to search is low, then displaced workers enter the uncovered sector, and the wage decreases.

The objective of this paper is to test Mincer's characterization of the effects of a minimum-wage hike on the uncovered sector. In the empirical work the covered sector is an aggregation of four low-wage nondurable manufacturing industries; the uncovered sector is agriculture prior to coverage under the Fair Labor Standards Act. Time-series data are used to estimate the covered-sector labor demand elasticity and the uncovered-sector wage and employment effects of a minimum-wage hike for nine regions of the United States. As theory predicts, in regions with comparatively small demand elasticities, the northern and midwestern regions, the uncovered-sector wage increases after a minimum-wage hike; and in regions with comparatively large demand elasticities, the southern and western regions, the uncovered-sector wage correspondingly decreases. The uncovered-sector employment effect is not as sharply estimated as the wage effect, and so its relationship to the demand elasticity is weak. This result is explained by limitations of the data.

There is a potentially serious problem with using regional time-series data to test the minimum-wage model. The problem arises because it is impossible to get long time series of data on the level of employment of persons earning exactly the minimum wage. Although average hourly earnings in the four low-wage industries are among the lowest in the economy, they are still well above the minimum wage, indicating that the data series contain contributions from high-wage labor. Thus the dependent variables in the regressions are total employment and the average wage, both aggregated over the wage distribution. It is conceivable, then, that the observed regional variation in the covered-sector demand elasticity is a statistical artifact reflecting regional differences in the mix of low- and high-wage labor. In other words, the true demand elasticity for minimum-wage grade labor could be the same in all regions, while the estimated demand elasticity is larger in the South simply because there are more low-wage workers there.

The evidence presented in subsequent sections suggests that the spurious aggregation effect is small. Briefly, the argument is as follows. Think of low- and high-wage labor as imperfect substitutes in a production function with constant returns to scale. Now let the minimum wage increase by 1 percent. There are both scale (output) and substitution effects. The scale effect causes employment of both factors, and hence total employment, to decline equiproportionately. The substitution effect, on the other hand, causes low-wage employment to decline and high-wage employment to increase. Intuitively, and as is shown in Appendix A, the observed proportionate increase

in the average wage ($\Delta\overline{W}/\overline{W}$) is no less than the share (s_1) of low-wage labor in total labor costs; if the substitution effect is zero, then $\Delta\overline{W}/\overline{W}$ equals s_1. The regional estimates of $\Delta\overline{W}/\overline{W}$ from the regressions turn out to be very close to regional estimates of s_1 derived from independent wage distribution data. This can happen only if the major impact of a minimum-wage hike is the scale effect and not the substitution effect. Since the scale effect dominates, the observed proportionate decline in total employment must be close to the proportionate decline in low-wage employment; that is, the aggregation bias is small in the covered sector.

The importance of the scale effect together with the cost-share data can explain why there are regional differences in the demand elasticity. The size of the scale effect depends upon the increase in the marginal cost of production caused by a hike in the minimum wage. Since s_1 is 0.14 in the North and 0.32 in the South, the proportionate increase in marginal cost is larger in the South, and so output and total employment contract proportionately more there. This is consistent with the idea that one purpose of the national minimum wage is to reduce the cost advantage of firms located in low-wage regions of the country. It is also consistent with Silberman and Durden's (1976, p. 325) finding that Southern congressmen are less likely than other congressmen to vote in favor of a minimum-wage hike.

The remainder of this paper is organized as follows. Section II describes a simple two-sector labor market model which is used to guide the empirical work. Sections III and IV report the empirical findings for the covered and uncovered sectors. Section V uses the theoretical model to interpret the results. Section VI contains the concluding remarks.

II. Two-Sector Model

The following two-sector model of the low-wage labor market is used to illustrate the restrictions being tested and to guide the empirical work. The model consists of two labor demand schedules (one for each sector), an equilibrium condition connecting the minimum wage and the competitively determined uncovered-sector wage, and an aggregate labor supply relationship. Formally, the model is:

$$\text{labor demand (covered):} \quad N^c = D^c(W/W_2, X^c), \quad D_1^c < 0, \tag{1a}$$

$$\text{labor demand (uncovered):} \quad N^u = D^u(W^u/W_2, X^u), \quad D_1^u < 0, \tag{1b}$$

$$\text{equilibrium condition:} \quad W^u = \left(\frac{\delta N^c}{U + \delta N^c}\right) W, \quad \delta > 0, \tag{1c}$$

$$\text{labor supply:} \quad N^c + N^u + U = S(W^u/P), \quad S' > 0, \tag{1d}$$

where N^c, N^u = employment in the covered (c) and uncovered (u) sector; W = minimum wage; W^u = wage in the uncovered sector; W_2 = control wage, $W_2 > W$; X^c, X^u = vectors of exogenous shift variables; U = unemployment; δ = turnover parameter; and P = general price level. The labor demand equation (1a) expresses employment in the covered sector, N^c, as a function of the relative minimum wage, W/W_2, where W_2 is an exogenously determined measure of high-skill wages, and as a function of exogenous shift variables, X^c. Similarly, equation (1b) expresses employment in the uncovered sector as a function of the uncovered-sector wage relative to the high-skill wage, W^u/W_2, and as a function of exogenous variables, X^u. Equation (1c) is from Mincer (1976). In equilibrium the uncovered-sector wage must equal the expected return from search in the covered sector (equal to the probability of getting a job times minimum wage). The parameter δ in (1c) is the proportion of covered-sector employees who each period must enter an employment lottery with the unemployed. Finally, equation (1d) relates the total labor force to the real wage in the uncovered sector.

The four endogenous variables in the system (1) are N^c, N^u, W^u, and U; the exogenous variables are W, W_2, P, X^c, and X^u. Note first that (1a) is the reduced-form expression for N^c. Now write the reduced form for the other three endogenous variables as

$$N^u = N^u(W, W_2, P, X^c, X^u), \tag{2a}$$

$$W^u = W^u(W, W_2, P, X^c, X^u), \tag{2b}$$

$$U = U(W, W_2, P, X^c, X^u). \tag{2c}$$

Theory places restrictions on the reduced-form equations. Specifically, Mincer shows

$$\frac{\partial W^u}{\partial W} \gtreqless 0 \quad \text{as} \quad e^c \lesseqgtr \delta, \tag{3a}$$

$$\frac{\partial N^u}{\partial W} \lesseqgtr 0 \quad \text{as} \quad e^c \gtreqless \delta, \tag{3b}$$

where e^c is the absolute elasticity of labor demand in the covered sector (the absolute elasticity of [1a] with respect to the minimum wage). The unemployment effect is positive,

$$\frac{\partial U}{\partial W} > 0, \tag{4}$$

and the sign of the labor force effect agrees with the sign of the uncovered-sector wage effect,

$$\frac{\partial N^c}{\partial W} + \frac{\partial N^u}{\partial W} + \frac{\partial U}{\partial W} \gtreqless 0 \quad \text{as} \quad e^c \lesseqgtr \delta. \tag{5}$$

Some of these results have been previously investigated empirically. There are, of course, many studies of the disemployment effect (1a) and of the unemployment effect (4). The Mincer study measures the labor force effect (5) for teenagers and young adults. After finding a negative labor force effect, Mincer infers $e^c > \delta$ and uses (3) to conclude that $\partial W^u/\partial W < 0$, that is, that, as a group, low-wage workers' prospects decline after a minimum-wage hike. Gardner's (1972) earlier work with aggregate U.S. agricultural data is consistent with Mincer's findings, though Gardner's estimates lack statistical significance at conventional levels. The purpose of this study is to use regional variation in the demand elasticity e^c to test the characterization (3a) and (3b). If there is small regional variation in the parameter δ, then regions with relatively small values of e^c should show $\partial W^u/\partial W > 0, \partial N^u/\partial W < 0$, and vice versa for regions with relatively large values of e^c.

III. The Covered Sector

The first step in the empirical work is to estimate the demand equation (1a) to get regional estimates of the covered-sector demand elasticity, e^c.[1] The data are annual observations on employment (1949–74) and average hourly earnings (1958–74) in an aggregation of four low-wage nondurable manufacturing industries: tobacco (SIC 21), textiles (SIC 22), apparel (SIC 23), and leather (SIC 31).[2] As mentioned in the Introduction, there is a problem because the employment variable includes workers for whom the legal minimum wage is economically ineffective. The effects of this skill-group aggregation can be characterized. Appendix A shows that regardless of the substitution pattern within aggregate labor, the observed (absolute) disemployment elasticity underestimates the sought-after demand

[1] The states within each region are: New England. Vt., Mass., N.H., Conn., R.I., Maine; Middle Atlantic: N.Y., Pa., N.J.; East North Central: Ohio, Wis., Ill., Ind., Mich.; West North Central: Mo., Iowa, N.Dak., S.Dak., Nebr., Kans., Minn.; South Atlantic: Va., Del., Ga., Md., W.Va., N.C., S.C., Fla.; East South Central: Ky., Tenn., Miss., Ala.; West South Central: Ark., La., Tex., Okla.; Mountain: Nev., Wyo., Mont., N.Mex., Utah, Idaho, Colo., Ariz.; Pacific: Calif., Oreg., Wash.

[2] Two-digit manufacturing industries were defined as low-wage industries if U.S. average hourly earnings in 1960 were no more than $1.70; the legal minimum was upped from $1.00 to $1.15 in September 1961. Only the four indicated nondurable manufacturing industries met the criterion. No other two-digit nondurable manufacturing industry had an average wage below $2.00 in 1960; the lowest average wage in 1960 of two-digit durable industries was $1.88 in furniture (SIC 25). Annual observations on state employment and average hourly earnings were collected from U.S. Bureau of Labor Statistics (1977). In the case of missing values, the data from nearby larger states were spliced into the shorter series. This allowed development of usable state employment series from 1949 forward and wage series from 1958 forward. The series were aggregated over the four industries and over contiguous states to nine regions.

TABLE 1

COVERED-SECTOR EMPLOYMENT AND WAGE EQUATIONS: SIGNIFICANCE TESTS[a]

	Dependent Variable: n^c (1949–74) Significance Test $F(3,17)$		Dependent Variable: w^c (1958–74) Significance Test $F(3,8)$	
	w [−1 to +1]	w_2 [−1 to +1]	w [−1 to +1]	w_2 [−1 to +1]
United States	****	****	*	****
New England	..	****	...	****
Middle Atlantic	...	****	...	****
East North Central	****	****	...	****
West North Central	*	****	**	****
South Atlantic	****	****	**	****
East South Central	****	****	***	****
West South Central	****	****	****	****
Mountain	****	****	N.A.[b]	N.A.
Pacific	****	****	...	****

NOTE.—Dependent variables. n^c = *log of total employment, four low-wage nondurable manufacturing industries,* by region as indicated; w^c = log of average hourly earnings, four low-wage nondurable manufacturing industries, by region as indicated. Independent variables. w = log of minimum wage, w_2 = log of average hourly earnings, all manufacturing, excluding average hourly earnings in the four low-wage industries, z = log of the FRB index of industrial production, all manufacturing. Lags are in brackets, negative lags are leads (future terms).

[a] Test that indicated group of distributed lag coefficients is zero.

[b] No wage data are available for the Mountain region.

* 15%

** 10%

*** 5%

**** 1%

elasticity. In addition, the observed elasticity of the covered-sector wage with respect to the minimum wage overestimates the share of minimum-wage grade labor in total labor costs. Below, evidence will be presented that e^c is only slightly underestimated.

To focus attention on economics the less important technical details of the regressions are put in Appendix B. The results in tables 1 and 2 are from distributed lag regressions of the logs of covered-sector employment and average hourly earnings, n^c and w^c, on the log of the minimum wage, w, the control wage, w_2, and the cyclical index, z. The control wage is the log of average hourly earnings, all manufacturing, net of wages in the four low-wage industries; the cyclical variable is the log of the FRB index of industrial production.[3] Each regression

[3] w, w_2, and z are the logarithms of the annual average of monthly values. The level W_2 was computed as $W_2 = (EW - \Sigma_j E_j W_j)/(E - \Sigma_j E_j)$, where E is production worker employment and W is production worker average hourly earnings. Unsubscripted values refer to all manufacturing; the subscript j indexes the four low-wage industries. The basic data were taken from U.S. Bureau of Labor Statistics (1976) and from issues of *Employment and Earnings* (U.S. Bureau of Labor Statistics 1976–77). The industrial production index was taken from U.S. Department of Commerce (1976) and issues of the *Survey of Current Business* (U.S. Department of Commerce 1976–77).

TABLE 2

COVERED-SECTOR EMPLOYMENT AND WAGE EQUATIONS: SUMS OF DISTRIBUTED LAG COEFFICIENTS
(Long-Run Effects)

	DEPENDENT VARIABLE: n^c			DEPENDENT VARIABLE: w^c		
	w [−1 to +1]	w_2 [−1 to +1]	z [0]	w [−1 to +1]	w_2 [−1 to +1]	z [0]
United States	−.2897	−.3124	.4370	.1865	.6936	−.0164
	(.0328)	(.0599)	(.0373)	(.0872)	(.1205)	(.0591)
New England	−.0394	−.9013	.5185	.1574	.5848	−.0836
	(.0847)	(.1468)	(.0794)	(.1148)	(.1590)	(.0777)
Middle Atlantic	−.1242	−.5710	.4159	.1406	.5827	−.1322
	(.0545)	(.0856)	(.0508)	(.1045)	(.1410)	(.0704)
East North Central	−.3812	−.3972	.5066	−.0099	.4943	−.2096
	(.0714)	(.1209)	(.0671)	(.1289)	(.1790)	(.0872)
West North Central	−.0987	.4048	.4261	.3500	.2371	−.2633
	(.0712)	(.0997)	(.0640)	(.1134)	(.1541)	(.0766)
South Atlantic	−.2237	−.2899	.4270	.1580	.8821	.1406
	(.0571)	(.0925)	(.0533)	(.0725)	(.0989)	(.0490)
East South Central	−.2042	−.2431	.5651	.3569	.6847	.0319
	(.0496)	(.0687)	(.0443)	(.0713)	(.0926)	(.0471)
West South Central	−.6360	−.0883	.6708	.4442	.6483	.0340
	(.0875)	(.1185)	(.0774)	(.1456)	(.1197)	(.0985)
Mountain	−.7589	.5059	1.4860	N.A.	N.A.	N.A.
	(.3909)	(.5911)	(.3606)			
Pacific	−.3984	1.0035	.5304	.2072	.4694	−.0685
	(.0746)	(.1178)	(.0694)	(.1366)	(.1899)	(.0924)

NOTE.—SE in parentheses. In the employment and wage equations df = 17, 8, respectively. See table 1 for the definitions of the variables. N.A. = data not available. Lags are in brackets.

included trend and a constant.[4] The regressions do not control for state minimum wages because in this sample state minimums are either at or below the federal minimum wage.[5]

The F-tests in table 1 indicate that the distributed lag coefficients on the minimum wage are significant as a group in nearly all of the equations. Now consider the long-run elasticities reported in table 2. As theory predicts, the sign of the minimum-wage effect is negative in all of the employment equations and positive in all but one wage equation. Moreover, the magnitude of the effect displays important regional variation. The disemployment effect appears to be negligible in the Northeast and upper Midwest, and it seems to be rather important through the South, the Southwest, and the Pacific coast. Also, the elasticities of employment and the average wage with respect to the minimum wage are consistent with one another, in the sense that, wherever the effect on the wage is small, so is the effect on employment.

To my knowledge, there are no other comparable published estimates of the disemployment effect by region or state. The estimated elasticity of −0.29 for the United States, however, is comparable with the findings of many other researchers. It is just outside Gramlich's (1976, p. 431) bounds of −0.05 to −0.25 for published estimates.[6]

Comparing these long-run effects to regional data on the wage distribution in the covered sector suggests that the major impact of the minimum wage is the scale or output effect. To see this, consider the 1964 wage distribution displayed in table 3. These data were

[4] Throughout this paper regional dependent variables are regressed on national variables. This is due to data limitations, but a strong case can be made for this practice even if regional data were available. Truly exogenous variables are driven by broad economy-wide forces and should not contain endogenous local fluctuations. For the quarterly agricultural regressions of Section IV it was possible to compute regional agricultural prices but only on an annual basis. Quarterly models with these prices repeated four times gave much lower $\bar{R}^2$ than the reported equations. Evidently, the time variation is more important than the regional variation.

[5] In the Northeast, state minimum wages tend to be just at the federal wage, and outside the Northeast they are usually well below the federal wage. Movements in state wage floors tend to parallel changes in the federal wage, and the main effect of state minimums is to extend coverage to industries only partly covered by federal legislation: retail trade, laundries, services, etc., industries excluded from the sample. In 1970 the federal minimum was $1.60, while the simple averages of state minimums by region were: New England, $1.60; Middle Atlantic, $1.60; East North Central, $.98; West North Central, $.65; South Atlantic, $.77; East South Central, $.18; West South Central, $.84; Mountain, $1.07; and Pacific, $1.50. New York was the only state with a minimum wage above $1.60; its minimum was raised from $1.60 to $1.85 effective July 1, 1970. Washington, D.C., and Alaska had 1970 minimums of $1.80 and $2.00, respectively, but were excluded from the sample. Source: U.S. Employment Standards Administration (1972*b*, p. 52, table 5).

[6] Many of these studies use the ratio of teenage to adult employment as the dependent variable, and so they estimate only a pure substitution effect (Welch 1974). This study estimates a combined scale and substitution effect.

TABLE 3

WAGE DISTRIBUTION FOR THE AGGREGATION OF FOUR NONDURABLE MANUFACTURING INDUSTRIES, March 1964 (%)

Straight-Time Wage ($) Less Than	United States	Northeast	South	North Central	West
1.25	1	1	1	1	1
1.30	16	11	22	18	14
1.35	23	16	30	23	19
1.40	29	21	39	31	22
1.45	36	28	47	37	28
1.50	41	32	53	42	31
2.50	91	85	98	94	83
3.50	98	96	100	100	97
Total	100	100	100	100	100
Average hourly earnings ($)	1.76	1.90	1.59	1.72	1.92
Under $1.40 shares of total:					
Employment (k_1)	.29	.21	.39	.31	.22
Labor costs (s_1)	.22	.14	.32	.24	.15

NOTE.—In September 1963 the basic minimum wage was raised from $1.15 to $1.25, in February 1967 it was raised to $1.40. The distributions were computed from table 4 of U.S. Wage and Hour and Public Contracts Divisions (1965). Accounting identities and interpolations were used to get complete data for the North Central and West regions, and so the distributions for these two regions are less reliable than the other distributions.

collected in a special survey of firms by the Department of Labor. Define low-wage workers to be those within 10 percent of the minimum wage of $1.25 (10 percent was about the average percentage increase in the minimum wage in the early 1960s). Then the share of low-wage labor in total labor costs is 0.14 in the Northeast and 0.32 in the South. These cost shares indicate that, if firms held employment constant, then a 10 percent minimum-wage hike would drive up labor costs by 1.4 and 3.2 percent in these two regions. Furthermore, if firms respond to the wage hike by contracting low- and high-wage employment equiproportionately, then the average wages in the North and South would increase by 1.4 and 3.2 percent. According to the regressions in table 2, the average wages actually increase by 1.5 percent in the North and 3.2 percent in the South. (These figures are 10 times the unweighted mean elasticity for the two northern regions and three southern regions.) A similar agreement between the cost share and the wage elasticity exists in the North Central region and in the West. The data appear to say that covered-sector firms respond to a minimum-wage hike by taking steps to reduce total labor input and that both groups of workers suffer equiproportionate disemployment. Furthermore, since the substitution effect appears to be small

relative to the scale effect, the estimated disemployment elasticities are close to the true demand elasticities for low-wage labor.

IV. The Uncovered Sector

The purpose of this section is to estimate the minimum-wage effect on uncovered (agricultural) employment and wages. The reduced-form equations (2a) and (2b) give the specifications. Specifically, the logs of agricultural employment and wages are regressed on the minimum wage, the control wage, the general price level, and the shift variables appearing in (1a) and (1b). Data are available quarterly, and the operational definitions are as follows. Employment and wages, n^u and w^u, are, respectively, the logs of hired farm employment and the hourly wage exclusive of room and board. Agricultural economists consider this wage to be more reliable than the U.S. Department of Agriculture's various composite measures. In the list of independent variables, the minimum wage variable is cw, the legal minimum multiplied (before logging) by the Moore index of the proportion of nonagricultural employees covered by the legislation. Intuitively, the impact of a 10 percent minimum-wage hike on the uncovered agricultural labor market is much different if 20 versus 90 percent of the nonagricultural work force is covered. The control wage w_2 and the cyclical index z are the same variables as used in Section III. The general price level p is the log of the CPI. Finally, the uncovered-sector shift variables are the logs of the USDA's index of prices received by farmers, p^u, and the index of prices paid by farmers for production items q. The index q excludes wages.[7]

The basic sample period (before lags and leads) is 1947:I–1966:IV. Truncation at 1966 restricts the estimation to a period in which agriculture was effectively uncovered by federal legislation. On February 1, 1967, coverage was extended to employees of large farms. The Department of Labor estimated that by 1971 about 48 percent of the agricultural work force was covered by federal legislation.[8] Also, after extension of federal coverage to agriculture, several states expanded their coverage to agriculture. Prior to 1967, state agricultural minimums were considered unimportant.[9]

[7] The agricultural employment and wage series were obtained from various issues of *Farm Labor* (U.S Department of Agriculture 1951–75). The input and output prices indexes were obtained from issues of *Agriculture Prices, Annual Summary* (U.S. Department of Agriculture 1961–72).

[8] For details see U.S. Employment Standards Administration (1972*a*, pp. 5–9).

[9] In 1966 only three states in the sample had a minimum wage applicable to agricultural workers: California, $1.30; Michigan, $1.15; and Wisconsin, $1.00. The California and Wisconsin minimums applied only to women and minors. Source: U.S. Wage and Hour and Public Contracts Divisions (1966, p. 30, table 12). Michigan and Wiscon-

Two conclusions are apparent from the regressions. First, the significance tests in table 4 confirm the well-known fact that aggregate cyclical forces play a strong role in the determination of agricultural labor supply. According to the U.S. long-run elasticities in table 5, a 10 percent increase in industrial production is associated with an 11.3 percent decrease in hired farm employment and a 2.9 percent increase in the hired farm wage. This is important because it indicates that there are indeed sizable labor flows between agriculture and the covered sector.

Second, and most important, changes in the minimum wage do affect agricultural employment and wages. The F-tests in table 4 show that as a group the distributed lag coefficients on the minimum wage are significant at 10 percent in five of 10 employment equations and in eight of 10 wage equations. Note that in table 5 the long-run U.S. employment elasticity of $-.18$ is insignificant, but the wage elasticity of $-.08$ is significant. The latter measurement indicates that for the entire United States a minimum-wage hike tends to depress the uncovered-sector wage. This finding is consistent with both Mincer's and Gardner's results.

Examination of the regional results in table 5 suggests that there are important regional differences in the welfare effects of the minimum wage. Notice that the effect on the uncovered-sector wage is significantly positive in the New England, Middle Atlantic, and West North Central regions; it is significantly negative in the South Atlantic and Pacific regions. Thus, as indicated by (3a), the wage in the uncovered sector can indeed rise or fall after a minimum-wage hike in the covered sector. Viewed in isolation, however, the uncovered-sector results are insufficient to verify the model. Indeed, if the hypothesized economic mechanisms are really at work, then the absolute disemployment elasticities from Section III should be largest in the regions where the uncovered-sector wage falls.

V. Sectoral Comparisons

Consider the inequality (3a). If the demand elasticity e^c is less than the turnover parameter δ, then after a minimum-wage hike the flow of labor out of the uncovered sector increases the wage there; if e^c is greater than δ, then the flow of labor into the uncovered sector decreases the wage. Therefore, if there is small regional variation in

sin are in the East North Central region, and their agricultural minimum wages possibly could explain the results for that region. Consider, though, the fact that California accounts for 88 percent of agricultural employment in the Pacific region, and it turns out that after a hike in the U.S. minimum wage the agricultural wage *declines* in the Pacific region.

TABLE 4

UNCOVERED-SECTOR EMPLOYMENT AND WAGE EQUATIONS: SIGNIFICANCE TESTS

	Dependent Variable n^u (1948:IV–1966:I)						Dependent Variable w^u (1948:IV–1966:I)					
	cw [0–10]	w_2 [0–5]	z [0–4]	p^u [0–5]	p [0–7]	q [0–7]	cw [0–10]	w_2 [0–5]	z [0–4]	p^u [0–5]	p [0–7]	q [0–7]
United States	***	****	***	***	*	****	****	****	****	****	****	****
New England	...	...	..	...	..	..	****	****	****	****	****	****
Middle Atlantic	..	...	*	...	...	...	****	**	****	****	**	***
East North Central	...	...	***	***	*	*	**	...	****	***	**	...
West North Central	**	****	*	***	..	***	...	...	****	****	...	****
South Atlantic	**	***	***	***	**	***	****	*	****	***	***	...
East South Central	**	*	**	...	*	...	****	...	****	***	****	**
West South Central	...	**	****	...	..	***	...	...	...	...	...	...
Mountain	**	****	**	*	...	*	***	*	...	*	**	*
Pacific	*	***	***	...	...	...	****	***	****	****	***	****

NOTE.—Test statistics are F-statistics with numerator degrees of freedom equal to the number of included lag coefficients for each variable and denominator degrees of freedom equal to 18. Lags are in brackets. Dependent variables: n^u = log of hired farm employment, by region as indicated; w^u = log of hourly wage for hired farm labor exclusive of room and board, by region as indicated. Independent variables: cw = log of coverage-weighted minimum wage; w_2 = log of average hourly earnings, all manufacturing excluding average hourly earnings in the four low-wage manufacturing industries; z = log of the FRB index of industrial production, all manufacturing; p^u = log of the USDA index of prices received by farmers; p = log of the consumer price index, all items; q = log of the USDA index of prices paid by farmers for production items.

* 15%.

** 10%.

*** 5%.

**** 1%.

TABLE 5

UNCOVERED SECTOR EMPLOYMENT AND WAGE EQUATIONS: SUMS OF DISTRIBUTED LAG COEFFICIENTS (Long-Run Effects)

	Dependent Variable n^u						Dependent Variable w^u					
	cw [0–10]	w_z [0–5]	z [0–4]	p^u [0–5]	p [0–7]	q [0–7]	cw [0–10]	w_z [0–5]	z [0–4]	p^u [0–5]	p [0–7]	q [0–7]
United States	−.18	3.23	−1.13	1.29	−3.60	−.60	−.08	1.28	.29	.71	−1.59	−.08
	(.13)	(1.44)	(.29)	(.85)	(2.89)	(1.29)	(.02)	(.23)	(.07)	(.14)	(.47)	(.22)
New England	−.17	5.99	−1.24	−.97	−3.60	1.57	.07	1.12	.29	.78	−.92	−.82
	(.39)	(4.40)	(.90)	(2.61)	(8.77)	(3.94)	(.03)	(.35)	(.09)	(.20)	(.71)	(.30)
Middle Atlantic	.09	.87	−.95	.14	1.33	−.13	.08	.63	.26	.15	.22	−.23
	(.19)	(2.32)	(.49)	(1.30)	(4.72)	(1.83)	(.02)	(.23)	(.07)	(.13)	(.48)	(.21)
East North Central	−.15	5.82	−1.17	3.28	−7.49	−2.80	.05	−.14	.56	−.11	2.14	−.30
	(.24)	(2.54)	(.50)	(1.48)	(5.22)	(2.13)	(.04)	(.43)	(.09)	(.24)	(.89)	(.36)
West North Central	−.27	7.58	−.85	4.83	−9.85	−3.95	.06	.73	.34	.98	.41	−1.25
	(.18)	(1.99)	(.41)	(1.14)	(4.02)	(1.63)	(.03)	(.37)	(.08)	(.20)	(.77)	(.28)
South Atlantic	.09	2.10	−1.12	1.55	−5.79	−1.12	−.10	−1.19	.79	−.53	3.55	−.17
	(.19)	(1.95)	(.42)	(1.13)	(3.98)	(1.63)	(.05)	(.56)	(.12)	(.30)	(1.21)	(.41)
East South Central	.29	−1.55	−.00	−1.07	5.07	1.78	.08	1.42	.24	.88	−2.57	−.34
	(.45)	(4.60)	(.87)	(3.41)	(8.95)	(5.26)	(.08)	(.84)	(.18)	(.55)	(1.64)	(.83)
West South Central	−.57	2.55	−2.26	1.62	1.87	−2.70	−.05	.06	.50	.19	−.91	.58
	(.31)	(4.40)	(.68)	(2.25)	(9.46)	(3.67)	(.04)	(.47)	(.11)	(.29)	(.94)	(.44)
Mountain	−.11	5.83	−.07	1.89	−9.76	−.77	−.07	.62	.13	1.00	−2.07	−.57
	(.17)	(1.73)	(.37)	(.98)	(3.60)	(1.42)	(.06)	(.70)	(.14)	(.50)	(1.38)	(.75)
Pacific	−.15	2.45	−1.14	.79	−1.48	−.33	−.12	.75	.49	.84	.13	−1.16
	(.16)	(1.70)	(.35)	(1.00)	(3.40)	(1.51)	(.03)	(.32)	(.08)	(.18)	(.66)	(.28)

NOTE.—SE in parentheses. In both equations, df = 18. See table 4 for definitions of variables. Lags are in brackets.

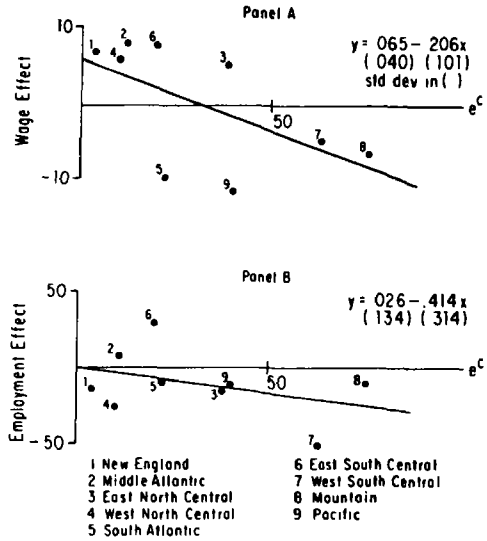


FIG. 1.—The relationship between the uncovered-sector wage and employment effects and the covered-sector disemployment elasticity.

the parameter δ, then plotting regional observations on the effect of the minimum wage on the wage in the uncovered sector against the demand elasticity e^c should give a downward sloping relationship crossing the horizontal axis at δ. By (3b), plotting the effect of the minimum wage on employment in the uncovered sector against e^c should give an upward sloping relationship also crossing the horizontal axis at δ. Panels *A* and *B* of figure 1 are the corresponding empirical diagrams. The dependent variable in panel *A* is the observed long-run elasticity of the uncovered-sector wage with respect to the minimum wage (from table 5); the dependent variable in panel *B* is the long-run uncovered-sector employment elasticity (also from table 5). The independent variable in each panel is the absolute value of the long-run covered-sector disemployment elasticity (from table 2).

As predicted by theory, panel *A* indicates an inverse relationship between the effect of the minimum wage on the uncovered-sector wage and the demand elasticity. The fitted line's horizontal intercept of 0.32 is an estimate of the critical demand elasticity (δ) that determines the sign of the wage effect in the uncovered sector. The estimate $\delta = 0.32$ says that in the steady state at least 32 percent of covered-sector employees compete with the unemployed for employment. In contrast, panel *B* suggests no relationship between the

effect of the minimum wage on uncovered employment and the demand elasticity. The data points are essentially random deviations about the horizontal axis. Note that the F-tests in table 4 indicate that the minimum wage affects uncovered employment. The sign and size of the long-run effect, however, are measured very imprecisely.

The skill-group aggregation in the uncovered sector data can account for the apparent conflict between theory and panel B of figure 1. When the low-skill wage changes in the uncovered sector, there are scale and substitution effects. As noted before and as proved in Appendix A, the larger is the substitution effect relative to the scale effect, the smaller is the observed (absolute) employment elasticity relative to the true proportionate change in low-wage labor. Evidently, the elasticity of substitution between low- and high-skill labor in agriculture is large enough to make this skill-group aggregation suppress the uncovered employment effect to an undetectable level. Notice that this is the reverse of the situation in the covered sector. As was noted in Section III, in the covered sector the scale effect dominates the substitution effect; in contrast, in the uncovered sector the substitution effect dominates the scale effect.

VI. Summary and Conclusion

This paper examines the covariation over regions in the effects of the minimum wage on wages and employment in the covered and uncovered sectors. Its major empirical finding is that the pattern of results is consistent with theory. Specifically, after a minimum-wage hike the uncovered-sector wage increases in the five regions with relatively small covered-sector demand elasticities: New England, Middle Atlantic, East North Central, West North Central, and East South Central; it decreases in the four regions with the relatively large covered-sector demand elasticities: South Atlantic, West South Central, Mountain, and Pacific. In addition, Mincer's and Gardner's earlier work suggesting that the uncovered-sector wage decreases in the entire United States is consistent with these findings, for their results are dominated by the four large regions (in terms of low-wage employment) in which the uncovered-sector wage falls. This paper's results have some bearing on the issue of welfare effects of the minimum wage. It seems fair to conclude that the minimum wage increases the well-being of all low-wage workers in the New England, Middle Atlantic, East North Central, West North Central, and East South Central regions. In the remaining regions, however, the minimum wage increases the well-being of only those workers having covered-sector employment.

Appendix A

The purpose of this Appendix is to characterize the effects of using total employment and the average wage as dependent variables in the regressions. Suppose low-skill class 1 labor (N_1) and high-skill class 2 labor (N_2) are imperfect substitutes in a production function with constant returns to scale. Let the class 1 wage (W_1) increase. Then

$$\tilde{N}_1 = (\text{scale effect} - s_2\sigma_{12})\tilde{W}_1,$$

$$\tilde{N}_2 = (\text{scale effect} + s_1\sigma_{12})\tilde{W}_1,$$

where $\tilde{}$ denotes proportionate change, s_1 and s_2 are the factors' shares in expenditures, and $\sigma_{12} > 0$ is the absolute Hicks-Allen partial elasticity of substitution. Let k_1 and k_2 denote the factors' shares in total employment; then the proportionate change in total employment is

$$\begin{aligned}\tilde{N} &= k_1\tilde{N}_1 + k_2\tilde{N}_2 \\ &= [\text{scale effect} + (k_2 - s_2)\sigma_{12}]\tilde{W}_1,\end{aligned}$$

which is less negative than $\tilde{N}_1$. The change in the average wage is

$$\begin{aligned}\tilde{\overline{W}} &= (k_1\tilde{W}_1 + k_2\tilde{W}_2) \\ &= s_1\tilde{W}_1 + (s_1 - k_1)(\tilde{N}_1 - \tilde{N}_2) \\ &= [s_1 - (s_1 - k_1)\sigma_{12}]\tilde{W}_1,\end{aligned}$$

which exceeds $s_1\tilde{W}_1$ because $s_1 < k_1$ (class 1 labor is the low-wage factor). Clearly the characterization applies equally well to either sector. (The above arguments were found by an anonymous referee.)

Appendix B

This Appendix discusses the technical econometric issues associated with the empirical work.

Specification

There are three important points concerning the covered-sector equations. First, adjustment costs (hiring and separation costs) induce a dependence of factor demand on lagged and expected future factor prices (Sargent 1978). This motivates using distributed lag methods and including a lead value of the minimum wage, for its movements are known well in advance. Second, minor distortions due to temporal aggregation (the unavoidable use of annual instead of monthly data) are apparently the cause of a significant lead coefficient on the control wage. This was verified by using Geweke's (1978) results. The effects of the temporal aggregation are minor because the explanatory variables are annual averages and the long-run effects are therefore estimated consistently (see Geweke's theorem 2). Third, the temporal aggregation and the forecasting of factor prices are good reasons to expect the homogeneity of degree zero of (1a) in nominal variables to break down in practice. Consequently, w and w_2 are entered separately in (1a) and the homogeneity constraint tested (and rejected). The coefficients have a "real" interpretation, for the entire history of each nominal variable is held constant.

The same comments apply to the uncovered sector regressions, though

TABLE B1

Exogeneity Tests

	Covered Sector								Uncovered Sector[a]											
	Dependant Variable n^c				Dependent Variable u^c				Dependent Variable n^u						Dependent Variable u^u					
	w (−2)	w_2 (−1)	w_2 (−2)	z (−1)	w (−2)	u_2 (−1)	w_2 (−2)	z (−1)	cw	w_2	z	p^u	p	q	cw	w_2	z	p^u	p	q
United States	.	****	. .	. . .	. . .	***	. .	. .	. .	. .	. . .	. . .	. .	. . .	. .	.	. . .	. .	. . .	. .
New England	. . .	. . .	. . .		. .	. .	. .	. .	. . .	. .	.	. .	. .	. .	. . .	. .	. . .	. .	. .	. .
Middle Atlantic	. .	.	.	. . .	. .	**	. .	. . .	. . .	. .	. . .	. . .	***	**	. . .	. . .	**	. .	. . .	. . .
East North Central	. . .	*	. . .	. .	. . .	. .	. . .	. .	.	. .	. .	. . .	.	**	***	****		. . .	. . .	. . .
West North Central	. .	. .	. . .	. .	. . .	. .	. . .	.	. . .	. . .	. .	. .	.	.	. .	. .	. . .	. . .	. . .	. . .
South Atlantic	. .	****	*	. . .	. .	***	. .	*	. .	. .	. .	. .	. . .	. .	. . .	. . .	. . .	***	. . .	. . .
East South Central	.	. . .	. .	. .	. .	***		.	.	.	. .	. . .	. .	. . .	. . .	. . .	. . .	. . .	. . .	. . .
West South Central	***	****	. .		. .	. . .	. .	. .	. .	. .	. . .	. .	. .	. . .	. .	. . .	. . .	. .	. .	. . .
Mountain	***	****	. . .	. .	N.A.	N.A.	N.A.	N.A.	.	**	. . .	. . .	. . .	. . .	*	. . .	. . .	. . .	. . .	. .
Pacific	. .	****	. .	. . .	. . .	.	. . .	. . .	. . .	. .	**	. .	. . .	. .	. .	. . .	**	***	. . .	. . .

Note.—Tests are for indicated lead coefficients equal to zero. The statistics are $t(17)$ in the n^c equation, $t(8)$ in the u^c equation, and $F(3,18)$ in both the n^u and u^u equations. Leads are in brackets.

[a] All leads in the n^u and u^u equations are −1 to −3.

* 15%

** 10%

*** 5%

**** 1%

with quarterly data the temporal aggregation problems are absent. To purge the series of seasonal effects, each regression included seasonal dummies and the products of seasonals with trend.

Serial Correlation

In the annual covered-sector regressions a first-order Cochrane-Orcutt correction was used; in the quarterly uncovered-sector regressions ad hoc prefilters were used together with fifth-order Cochrane-Orcutt.

Exogeneity

Table B1 reports the outcome of batteries of Sims's (1972, 1977) exogeneity test. The test consists of the appropriate F- or t-test on lead coefficients. The paucity of rejections indicates no serious problems with a lack of exogeneity in either set of results. Since lead values of w and w_2 enter the covered-sector equations for other reasons, the test's power is obviously low in that set of results.

References

Gardner, Bruce. "Minimum Wages and the Farm Labor Market." *American J. Agricultural Econ.* 54 (August 1972): 473–76.

Geweke, John F. "Temporal Aggregation in the Multiple Regression Model." *Econometrica* 46 (May 1978): 643–61.

Gramlich, Edward M. "Impact of Minimum Wages on Other Wages, Employment, and Family Incomes." *Brookings Papers Econ. Activity*, no. 2 (1976), pp. 409–51.

Mincer, Jacob. "Unemployment Effects of Minimum Wages." *J.P.E.* 84, no. 4, pt. 2 (August 1976): S87–S104.

Sargent, Thomas J. "Estimation of Dynamic Labor Demand Schedules under Rational Expectations." *J.P.E.* 86 (December 1978): 1009–44.

Silberman, Jonathan I., and Durden, Garey C. "Determining Legislative Preferences on the Minimum Wage: An Economic Approach." *J.P.E.* 84 (April 1976): 317–29.

Sims, Christopher A. "Money, Income, and Causality." *A.E.R.* 62 (September 1972): 540–52.

———. "Exogeneity and Causal Orderings in Macroeconomic Models." In *New Methods in Business Cycle Research: Proceedings from a Conference*, edited by Christopher A. Sims. Minneapolis: Federal Reserve Bank of Minneapolis, 1977.

U.S. Bureau of Labor Statistics. *Employment and Earnings, United States, 1909–75*. Statistical Bulletin no. 1312-10. Washington: Government Printing Office, 1976.

———. *Employment and Earnings*. Washington: Government Printing Office, 1976–77, various issues.

———. *Employment and Earnings, State and Areas, 1939–75*. Statistical Bulletin no. 1370-12. Washington: Government Printing Office, 1977.

U.S. Department of Agriculture, Statistical Reporting Service. *Agricultural Prices, Annual Summary*. Washington: Government Printing Office, 1961–72, various issues.

———. Statistical Reporting Service, Agricultural Marketing Service, Bureau

of Agricultural Economics. *Farm Labor.* Washington: Government Printing Office, 1951–75, various issues.

U.S. Department of Commerce. *Business Statistics.* 1975 ed. Washington: Government Printing Office, 1976.

———. *Survey of Current Business.* Washington: Government Printing Office, 1976–77, various issues.

U.S. Employment Standards Administration. *Hired Farm Workers.* Washington: Government Printing Office, 1972. (*a*)

———. *Wages and Hours of Nonsupervisory Employees in All Private Nonfarm Industries by Coverage Status under the Fair Labor Standards Act.* Washington: Government Printing Office, 1972. (*b*)

U.S. Wage and Hour and Public Contracts Divisions. *Manufacturing Industries: A Study to Evaluate the Minimum Wage and Maximum Hours Standards of the Fair Labor Standards Act.* Washington: Government Printing Office, 1965.

———. *Hired Farm Labor: A Study to Evaluate the Minimum Wage and Maximum Hours Standards of the Fair Labor Standards Act.* Washington: Government Printing Office, 1966.

Welch, Finis. "Minimum Wage Legislation in the United States." *Econ. Inquiry* 12 (September 1974): 285–318.

Taxes, Inflation, and the Durability of Capital

Andrew B. Abel

Harvard University

The choice of capital durability is affected by the rate of inflation because nominal depreciation deductions are based on historical cost rather than replacement cost. This paper analyzes a model with competitive firms and intertemporally optimizing consumers and demonstrates that an increase in the rate of inflation will lead to an increase in the durability of capital, if the ratio of the nominal discount rate to the rate of depreciation exceeds a critical value. However, if this ratio is less than the critical value, this effect is reversed. The effect of costs of adjustment on the critical value is analyzed.

I. Introduction

Since depreciation deductions for tax purposes are based on historical cost rather than replacement cost, inflation can affect the investment and capital durability decisions of firms. This problem was recently analyzed by Auerbach (1979), using a general equilibrium model in which investment and durability are determined by optimizing firms and consumption is determined according to a proportional savings function. To analyze the effect of inflation on the choice of asset durability, Auerbach simulates his model, using a Cobb-Douglas technology with reasonable parameter values.[1] His simulation results indicate that higher inflation leads firms to choose more durable capital, although we are left with a caveat about the robustness of the results.

This paper was written while I was an assistant professor at the University of Chicago Department of Economics. I thank Alan Auerbach, Dennis Carlton, Stanley Fischer, Robert LaLonde, Frederic Mishkin, and Michael Mussa for comments on earlier drafts of this paper.

[1] Auerbach also assumes values for the share of government expenditure in net output, the fraction of disposable income saved, and the corporate tax rate.

[*Journal of Political Economy*, 1981, vol 89, no 3]
© 1981 by The University of Chicago 0022-3808/81/8903-0001$01 50

By modifying Auerbach's model to make consumption decisions based on intertemporal utility maximization rather than on a proportional saving function, we can derive analytically the effect of inflation on durability without having to specify the form of the production function. The direction of the effect of inflation on durability is determined by whether the ratio of the nominal discount rate to the depreciation rate exceeds a critical value which depends solely on the corporate tax rate and the rate of the investment tax credit. If the ratio of the nominal discount rate to the depreciation rate exceeds the critical value—as is the case in Auerbach's simulations[2]—then an increase in the rate of inflation leads to an increase in durability. However, if the ratio of the nominal discount rate to the depreciation rate is less than the critical value, then an increase in the rate of inflation leads to less durable capital. This reversal of Auerbach's finding is more likely to apply to durable equipment rather than to structures since equipment has a higher depreciation rate than structures.[3]

A second modification of Auerbach's model is the incorporation of adjustment costs into the model. The presence of convex costs of adjustment removes the indeterminacy of the individual firm's rate of investment in the Auerbach model. It should be noted that one complication of introducing adjustment costs is that the critical value of the ratio of the nominal discount rate to the depreciation rate is no longer a function of only tax parameters. The critical value of this ratio will, in general, depend on technology, although we can place bounds on this critical value. Using these bounds on the critical value of the ratio of the nominal discount rate to the depreciation rate, we can still derive fairly strong results about the effect of inflation on durability based only on this ratio and on the tax parameters. In general, if the nominal discount rate exceeds the rate of depreciation, an increase in the rate of inflation reduces the rate of depreciation. If the nominal discount rate is "enough" less than the depreciation rate, where enough depends only on tax parameters, an increase in the rate of inflation increases the rate of depreciation. In the region between the upper and lower bounds on the critical value, the critical value depends on technology. Note, however, that this complication arises from the introduction of adjustment costs. In the absence of adjustment costs, the critical value of the ratio of the nominal discount rate to the depreciation rate depends only on tax parameters.

[2] For a more precise discussion of Auerbach's simulation results, see Section III

[3] This statement implicitly assumes that the investment tax credit rate is the same for structures and equipment, so that the critical value is the same for both types of capital. However, under the current U.S. tax laws, the investment tax credit is applicable to equipment but not to structures. Therefore, as will become clear in Section III, the critical value is smaller for equipment than for structures. This effect may weaken or reverse the statement in the text.

In Section II we develop a model of a competitive firm and analyze the investment and durability decisions. The model of the firm is incorporated into a general equilibrium model in Section III, where we analyze the effect of changes in the rate of inflation. Section IV presents concluding remarks.

II. The Model of the Firm

Since the model of the firm in this paper differs from Auerbach's model only by the inclusion of adjustment costs and the investment tax credit, the description of the model will be brief; further description may be found in Auerbach.

We assume that the firm is a price taker in its output market and in the labor market. The price of output p and the nominal wage rate w both inflate at rate π, so that the real wage is constant over time. Let δ_t be the (constant) rate of exponential decay of capital goods installed at time t. Ex ante, δ_t is a decision variable of the firm. We assume that the instantaneous flow of capital services from a unit of capital is an increasing concave function of δ, $A(\delta)$, with $A' > 0$, $A'' < 0$.

Investment in physical capital consists of purchasing a unit of the homogeneous output and installing it in place at some cost which is an increasing convex function of the rate of investment. Let I_t be nominal investment at time t, so that I_t/p_t is the number of units of output bought and installed as physical capital. We assume that the real cost of installation is an increasing convex function of real investment, so that the nominal cost of installation at time t is $p_t c(I_t/p_t)$, with $c(0) = c'(0) = 0$, and $c' > 0$, $c'' > 0$ for $I_t/p_t > 0$.[4] We assume that these installation costs can be expensed for tax purposes. Note that the assumption that the purchase of uninstalled capital goods is amortized over time while the installation cost is immediately expensed is similar to Auerbach's case in which a fraction e of the cost of investment is expensed.[5]

Capital services at time s are

$$KS_s = \int_{-\infty}^{s} A(\delta_t)(I_t/p_t)e^{-\delta_t(s-t)}dt \tag{1}$$

since $(I_t/p_t)e^{-\delta_t(s-t)}$ is the amount of capital installed at time t which remains at time s and $A(\delta_t)$ is the flow of capital services per unit of this

[4] Adjustment costs have been incorporated into models of investment by Eisner and Strotz (1963), Lucas (1967), Gould (1968), and Treadway (1969). Abel and Blanchard (1980) distinguish the cost of purchasing uninstalled capital from the cost of installing capital, as is done in the text.

[5] There is a slight difference, however, in that the fraction of the total cost which is expensed in this model is endogenous, whereas the fraction e is exogenous in Auerbach's model.

capital. Gross output at time s is given by $Y_s^G = H(KS_s, L_s)$, where H is first-degree homogeneous in KS and L. We also assume $H_{KS}, H_L, H_{KS,L} > 0$ and $H_{KS,KS}, H_{L,L} < 0$.

Taxable corporate profits, which are revenues less wages, installation costs, and depreciation allowances, are taxed at rate τ, $0 < \tau < 1$. Let $D(x, \delta)$ be the depreciation deduction per dollar of purchase cost of capital of age x which decays at rate δ, and let k $(0 \leq k < 1)$ be the investment tax credit rate applied to the purchase of uninstalled capital.

The objective of the firm is to maximize

$$V = \int_0^\infty \Big\{(1-\tau)[p_s Y_s^G - w_s L_s - p_s c(I_s/p_s)] - (1-k)I_s + \tau \int_{-\infty}^{s} I_v D(s-v, \delta_v)dv\Big\} e^{-rs} ds \tag{2}$$

with respect to L_s, δ_s, and I_s, where r is the rate at which equity holders discount after-tax nominal flows.[6] Note that (2) is the same as Auerbach's equation 3 except for the inclusion of installation costs and the investment tax credit. For simplicity of notation, let

$$z(\delta) = \int_0^\infty D(x, \delta)e^{-rx}dx \tag{3}$$

be the present value of depreciation deductions, and let $\rho = r - \pi$ be the after-tax real discount rate. Differentiating V with respect to δ_t, I_t, and L_t and setting these derivatives equal to zero, we obtain:[7]

$$(1-\tau)\int_t^\infty [A'(\delta_t) - (s-t)A(\delta_t)]H_{KS}(KS_s, L_s)e^{-(\rho+\delta_t)(s-t)}ds = \frac{\partial}{\partial \delta_t}[1 - k - \tau z(\delta_t)], \tag{4a}$$

$$(1-\tau)c'(I_t/p_t) = (1-\tau)A(\delta_t)\int_t^\infty H_{KS}(KS_s, L_s)e^{-(\rho+\delta_t)(s-t)}ds - [1 - k - \tau z(\delta_t)], \tag{4b}$$

$$H_L(KS_t, L_t) = w_t/p_t. \tag{4c}$$

These equations may be solved recursively.[8] At any point in time, t, KS_t is fixed and the firm chooses L_t to satisfy (4c). Note that, since the

[6] We assume that firms are financed entirely by equity, so that interest payments are not included in the expression for cash flow.

[7] The first-order conditions are identical to Auerbach's eq. 4, except that $(1-\tau)c'(I_t/p_t)$ in (4b) and k in (4a) and (4b) are zero in Auerbach's framework.

[8] If the installation costs were amortized over time rather than immediately expensed, then the system would be only block recursive. Eq. (4c) could be solved first, but the analogues of (4a) and (4b) would have to be solved simultaneously.

real wage, w_t/p_t, is assumed to be constant over time, $H_L(KS_t,L_t)$ is constant over time. Since $H(KS_t,L_t)$ is linearly homogeneous, the constancy of $H_L(KS_t,L_t)$ implies that $H_{KS}(KS_t,L_t)$ is constant over time. Let ϕ be the constant value of $H_{KS}(KS_t,L_t)$ and define $\lambda(\delta;\ \phi)$ as

$$\lambda(\delta_t;\ \phi) = \frac{(1-\tau)A(\delta_t)\phi}{\rho+\delta_t} + k + \tau z(\delta_t) - 1. \tag{5}$$

Note that $[(1-\tau)A(\delta_t)\phi]/(\rho+\delta_t)$ is the present value of the stream of after-tax marginal products accruing to the undepreciated portion of a unit of capital installed at time t. In addition, by installing a unit of capital at time t, the firm reduces the present value of its taxes by $k + \tau z(\delta_t)$. Therefore, $\lambda(\delta;\ \phi)$ is the excess of the present value of the after-tax cash flow accruing to the marginal unit of installed capital over the cost of an uninstalled unit of capital.

Note that the first-order conditions given in (4a) and (4b) may be expressed as[9]

$$\lambda'(\delta_t;\ \phi) = \frac{(1-\tau)\phi}{\rho+\delta_t}\left[A'(\delta_t) - \frac{A(\delta_t)}{\rho+\delta_t}\right] + \tau\frac{\partial z(\delta_t)}{\partial\delta_t} = 0, \tag{6a}$$

$$(1-\tau)c'(I_t/p_t) = \lambda(\delta_t;\ \phi). \tag{6b}$$

To summarize, the firm may use the following sequential decision process. Given a real wage rate, the firm chooses a KS/L ratio to equate H_L with the exogenous real wage. This KS/L ratio uniquely determines ϕ, the marginal product of capital services. Given ϕ, the excess of the shadow value of installed capital over the net cost of uninstalled capital is a function, $\lambda(\delta_t;\ \phi)$, of δ_t. According to (6a), the firm chooses δ_t to maximize $\lambda(\delta_t;\ \phi)$.[10] Given the resulting maximized value of $\lambda(\delta_t;\ \phi)$, the firm will choose the real rate of investment to equate the marginal installation cost with $\lambda(\delta_t;\ \phi)$, as in (6b). This solution procedure applies at all points in time, even when the capital stock has not yet reached its steady-state value. Since the real wage rate is assumed constant, the marginal product of capital services, ϕ, is also constant. Therefore, the λ-maximizing value of δ is constant over time, and the

[9] If we impose the constraint that gross investment is nonnegative, the first-order conditions with respect to I are: (*a*) $(1-\tau)c'(I_t/p_t) \geqq \lambda(\delta;\ \phi)$; and (*b*) $(I_t/p_t)[(1-\tau)c'(I_t/p_t) - \lambda(\delta_t;\ \phi)] = 0$. If $I_t/p_t > 0$, then (*b*) implies eq. (6b).

[10] If $z = \delta/(\rho+\delta)$, as will be the case later in this note, then $\lambda(\delta;\ \phi) = (k-1) + [(1-\tau)A(\delta)\phi/(\rho+\delta)] + [\tau\delta/(\rho+\delta)]$, $\lambda'(\delta;\ \phi) = [1/(\rho+\delta)^2]\{\phi(1-\tau)[A'(\delta)(\rho+\delta) - A(\delta)] + \tau\rho\}$, and $\lambda''(\delta;\ \phi) = [-2/(\rho+\delta)^3]\{(1-\tau)\phi[A'(\delta)(\rho+\delta) - A(\delta)] + \tau\rho\} + [(1-\tau)\phi A''(\delta)/(\rho+\delta)]$. Therefore, when $\lambda'(\delta;\ \phi)$ equals zero, the term in braces in the expression for $\lambda''(\delta;\ \phi)$ is zero and $\lambda''(\delta;\ \phi) = [(1-\tau)\phi A''(\delta)/(\rho+\delta)] < 0$. Hence, setting $\lambda'(\delta;\ \phi)$ equal to zero maximizes $\lambda(\delta;\ \phi)$ with respect to δ. Note that later, when we examine the case in which $z = [\delta/(\rho+\pi+\delta)]$, a similar line of reasoning will establish that $\lambda''(\delta;\ \phi) < 0$ when $\lambda'(\delta;\ \phi) = 0$, as long as π is not too large.

real rate of investment is also constant.[11] Given the constant value of real investment, say $(I/p)^*$, the steady-state capital stock is $(1/\delta)(I/p)^*$, and the steady-state flow of capital services is $[A(\delta)/\delta](I/p)^*$.

III. A General Equilibrium Model of Investment and Durability

In this section, we examine the effects of changes in the rate of inflation within the context of a general equilibrium model. Rather than posit a proportional savings function, as in Auerbach, we assume that consumption is determined as the outcome of infinite-horizon intertemporal utility maximization by identical households. As noted by Auerbach, the after-tax discount rate ρ will be constant in such a model.

To keep the analysis simple but without any essential loss of generality, assume that the number of households is constant, and each household owns one firm and inelastically supplies $\bar{L}$ units of labor to its firm. Each household selects c_t and m_t to maximize

$$\int_0^\infty U(c_t, m_t)e^{-\rho t}dt \tag{7}$$

subject to $c_t = y_t + (w_t/p_t)\bar{L} - (\dot{M}_t/p_t)$, where c_t is real consumption at time t, m_t is real balances at time t, $\dot{M}_t$ is the growth in nominal money balances, and $y_t \equiv (1-\tau)[H(KS_t,\bar{L}) - (w_t/p_t)\bar{L} - c(I_t/p_t)] - (1-k)I_t + \tau\int_{-\infty}^{t} I_s/p_t e^{-\pi(t-s)}D(t-s,\delta_s)ds$ is the after-tax real cash flow of the firm. The maximization problem is based on the assumption that each household receives the net cash flow of the firm in addition to wage income and uses its total income either to consume or to add to its money balances. We also assume that the government uses tax revenue and the issuance of new money to purchase current output—which does not affect the utility function $U(c_t, m_t)$.[12]

Solving the maximization problem (7), we can show that, in the steady state, equations (6a) and (6b) continue to hold. However, in this general equilibrium model with labor supply constant, ϕ, the marginal product of capital services, is a decreasing function of the steady-state level of capital services $[A(\delta)/\delta](I/p)$. Therefore, for the steady-state values of the variables of interest, ϕ, δ, and I/p, we have the following system of equations:

[11] Gould (1968) showed that, for a price-taking firm with constant returns to scale, the rate of investment is constant if the real wage rate is constant.

[12] Abel and Blanchard (1980) present a more complete discussion of a general equilibrium model with taxes and adjustment costs. However, in that model the depreciation rate is fixed exogenously, and money balances do not appear.

$$\phi = g\left[\frac{A(\delta)}{\delta} I/p\right]; \quad g' < 0, \tag{8a}$$

$$\lambda'(\delta, \phi) = 0, \tag{8b}$$

$$\lambda(\delta, \phi) - (1 - \tau)c'(I/p) = 0. \tag{8c}$$

Note that the system (8a)–(8c) could also represent the behavior of a single profit-maximizing firm with constant returns to scale with either some monopoly power in its output market or some monopsony power in the labor market.

To examine the effect of a change in the rate of inflation, we linearize the system (8a)–(8c) around its solution to obtain

$$\begin{bmatrix} 1 & -g' \cdot \frac{I/p}{\delta}\left[A'(\delta) - \frac{A(\delta)}{\delta}\right] & -g' \cdot \frac{A(\delta)}{\delta} \\ \lambda'_\phi & \lambda'_\delta & 0 \\ \lambda_\phi & 0 & -(1-\tau)c'' \end{bmatrix} \begin{bmatrix} d\phi \\ d\delta \\ d(I/p) \end{bmatrix} = \begin{bmatrix} 0 \\ -\lambda'_\pi \\ -\lambda_\pi \end{bmatrix} d\pi, \tag{9}$$

where $\lambda_\phi = [(1 - \tau)A(\delta)/(\rho + \delta)] > 0$, and $\lambda'_\phi = (1 - \tau)/(\rho + \delta)\{A'(\delta) - [A(\delta)/(\rho + \delta)]\} < 0$, and $\lambda'_\delta < 0$ because of the second-order condition for the optimal δ.[13] Let Δ be the determinant of the square matrix in (9), so that

$$\Delta = -\left\{\lambda'_\delta + \lambda'_\phi g' \cdot \frac{I/p}{\delta}\left[A'(\delta) - \frac{A(\delta)}{\delta}\right]\right\}(1 - \tau)c'' + g' \cdot \frac{A(\delta)}{\delta}\lambda_\phi\lambda'_\delta > 0. \tag{10}$$

Note that $\Delta > 0$ regardless of whether c'' is positive or zero.

Given the linearized system (9), Cramer's rule can be used to determine the effect on δ of changes in the rate of inflation π,

$$\frac{d\delta}{d\pi} = \frac{1}{\Delta}\left[(1 - \tau)c''\lambda'_\pi - g' \cdot \frac{A(\delta)}{\delta} B_\pi\right], \tag{11}$$

where $B_\pi = -\lambda'_\phi\lambda_\pi + \lambda_\phi\lambda'_\pi$. In analyzing (11), we will also consider the following two special cases: (*a*) the absence of adjustment costs, so that $c'' = 0$; and (*b*) the polar opposite case of extremely convex adjustment costs, so that $c'' = \infty$. In the case without adjustment costs, which corresponds most closely to Auerbach's model, it can be shown that

$$\frac{d\delta}{d\pi} = \frac{-1}{\Delta} g' \cdot \frac{A(\delta)}{\delta} B_\pi \quad (c'' = 0), \tag{12}$$

where the coefficient on B_π is positive. In the polar opposite case, it can be shown that

[13] Note that the element in the third row, second column of the square matrix in (9) is λ_δ, which is zero according to (8b).

$$\text{sign}\,\frac{d\delta}{d\pi} = \text{sign}\,\lambda'_\pi \quad (c'' = \infty). \tag{13}$$

In order to analyze the expressions for $d\delta/d\pi$, we must explicitly model the effect of inflation on the shadow price of installed capital. Following Auerbach, we assume that depreciation deductions reflect physical decay of capital but are based on historical cost, so that[14]

$$z(\delta) = \frac{\delta}{\rho + \pi + \delta}. \tag{14}$$

Using this specification of depreciation deductions, we can easily show that

$$\lambda = \frac{(1-\tau)A(\delta)\phi}{\rho + \delta} + k + \frac{\tau\delta}{\rho + \pi + \delta} - 1, \tag{15a}$$

$$\lambda' = \frac{(1-\tau)\phi}{\rho + \delta}\left[A'(\delta) - \frac{A(\delta)}{\rho + \delta}\right] + \frac{\tau(\rho + \pi)}{(\rho + \pi + \delta)^2}. \tag{15b}$$

Now we must calculate λ_π, λ'_π, and B_π. Differentiating λ with respect to π, we obtain

$$\lambda_\pi = \frac{-\tau\delta}{(\rho + \pi + \delta)^2} < 0. \tag{16}$$

Note that λ_π is negative because an increase in the rate of inflation reduces the present value of real depreciation deductions and hence reduces the shadow price of installed capital. Next, we differentiate λ' with respect to π to obtain

$$\lambda'_\pi = \frac{\tau}{(\rho + \pi + \delta)^3}(\delta - \rho - \pi) \gtrless 0 \quad \text{as } \delta \gtrless \rho + \pi. \tag{17}$$

Thus the sign of λ'_π depends on whether the rate of depreciation exceeds the nominal discount rate. As shown in figure 1, λ'_π is negative in region *I* and is positive in regions *II* and *III*.

Recall that $B_\pi = -\lambda'_\phi\lambda_\pi + \lambda_\phi\lambda'_\pi$, that λ_ϕ is positive, and that λ_π and λ'_ϕ are negative. Thus in region *I*, where $\delta < \rho + \pi$ and λ'_π is negative, B_π is negative. To determine the sign of B_π when $\delta > \rho + \pi$ requires more calculation. Using the expressions for λ_ϕ, λ'_ϕ, λ_π, and λ'_π, we can show that

$$B_\pi = \frac{1-\tau}{\rho + \delta}\,\frac{\tau A(\delta)}{(\rho + \pi + \delta)^2}\left\{\delta\left[\frac{A'(\delta)}{A(\delta)} - \frac{1}{\rho + \delta}\right] + \frac{\delta - \rho - \pi}{\rho + \pi + \delta}\right\}. \tag{18}$$

[14] If depreciation deductions are based on replacement cost rather than historical cost, then $z(\delta) = \delta/(\rho + \delta)$. In this case, changes in the rate of inflation have no effect on the firm's decision problem.

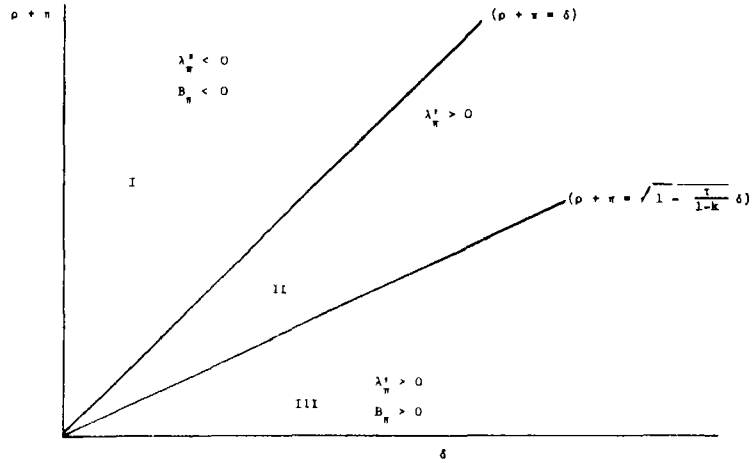


FIG. 1

It will now be useful to define q, the net cost of installed capital, ignoring the depreciation allowance, as

$$q = 1 - k + (1 - \tau)c'. \tag{19}$$

Using this expression for q, we obtain[15]

$$B_\pi = \frac{1 - \tau}{\rho + \delta} \frac{\tau A(\delta)}{(\rho + \pi + \delta)^3} \frac{q}{q(\rho + \pi + \delta) - \tau\delta} \left[\left(1 - \frac{\tau}{q}\right)\delta^2 - (\rho + \pi)^2\right]. \tag{20}$$

Note that B_π has the same sign as the term in brackets, so that

$$B_\pi \gtrless 0 \text{ as } \frac{\rho + \pi}{\delta} \lessgtr \sqrt{1 - \frac{\tau}{q}}. \tag{21}$$

Thus the sign of B_π depends on whether $(\rho + \pi)/\delta$ exceeds the critical value $\sqrt{1 - (\tau/q)}$. Note that the critical value $\sqrt{1 - (\tau/q)}$ is less than or

[15] Rewrite eqq. (6a) and (6b) as

$$\frac{(1 - \tau)\phi}{\rho + \delta}\left[A'(\delta) - \frac{A(\delta)}{\rho + \delta}\right] = \frac{-\tau(\rho + \pi)}{(\rho + \pi + \delta)^2}, \tag{6a'}$$

and

$$\frac{(1 - \tau)\phi}{\rho + \delta} A(\delta) = q - \frac{\tau\delta}{\rho + \pi + \delta}. \tag{6b'}$$

Dividing (6a') by (6b'), we obtain

$$\frac{A'(\delta)}{A(\delta)} - \frac{1}{\rho + \delta} = \frac{-\tau(\rho + \pi)}{q(\rho + \pi + \delta)^2 - \tau\delta(\rho + \pi + \delta)},$$

which may be substituted into (18).

TABLE 1

THE EFFECT OF INFLATION ON DURABILITY

	Sign of $\frac{d\delta}{d\pi}$		
Region	*I* $\frac{\rho+\pi}{\delta} > 1$	*II* $\sqrt{1-\frac{\tau}{1-k}} < \frac{\rho+\pi}{\delta} < 1$	*III* $\frac{\rho+\pi}{\delta} < \sqrt{1-\frac{\tau}{1-k}}$
General $c'' \geqslant 0$	−	?	+
No adjustment costs	−	−	+
$c'' = \infty$	−	+	+

equal to one, so that if $\rho + \pi > \delta$, as in region *I*, then $(\rho + \pi)/\delta$ exceeds the critical value, and B_π is negative. Also note that, since $q \geqslant 1 - k$, with equality holding in the absence of adjustment costs, the critical value must be greater than or equal to $\sqrt{1 - [\tau/(1 - k)]}$. Therefore, if $(\rho + \pi)/\delta < \sqrt{1 - [\tau/(1 - k)]}$ as in region *III*, B_π is positive.

Recall from (11) that, in the general case, if λ'_π and B_π have the same sign, then $d\delta/d\pi$ also has that sign. Thus, as summarized in table 1, $d\delta/d\pi$ is negative in region *I* and is positive in region *III*. In region *II*, λ'_π is positive, but the sign of B_π is indeterminate without further assumptions. If we assume that $c'' = \infty$, then the sign of $d\delta/d\pi$ is the same as the sign of λ'_π and hence is positive. However, if we assume that there are no adjustment costs, then the critical value is $\sqrt{1 - [\tau/(1 - k)]}$, and B_π is negative. Hence, in the absence of adjustment costs, $d\delta/d\pi$ is negative in region *II*.

To summarize, if the nominal discount rate exceeds the rate of depreciation, then an increase in the rate of inflation reduces the optimal δ and hence leads to more durable capital. On the other hand, if the nominal discount rate is less than $\sqrt{1 - [\tau/(1 - k)]}$ times the depreciation rate, an increase in the rate of inflation leads to less durable capital. In the case in which there are no adjustment costs and the investment tax credit is zero (as in Auerbach), an increase in the rate of inflation increases the durability of capital if $\rho + \pi > \sqrt{1 - \tau}\,\delta$ but reduces the durability of capital if $\rho + \pi < \sqrt{1 - \tau}\,\delta$.

It is interesting to compare the results for the case with zero adjustment cost and zero investment tax credit with the simulation results in Auerbach. Note that these two models are not identical in that they treat consumption decisions and money-holding decisions differently. The essential difference is that, in the optimizing model of this paper, ρ is held constant, whereas ρ is a variable in the Auerbach model. In Auerbach's table 1, the nominal discount rate exceeds the depreciation rate in every row except the first. Consistent

with the results derived above, an increase in the rate of inflation increases the durability of capital in Auerbach's simulations.

In the first row of Auerbach's table 1, the nominal discount rate falls short of $\sqrt{1-\tau}\,\delta$, and the results derived above would indicate that $d\delta/d\pi$ is positive. However, δ decreases when π is increased from row 1 to row 2. There are two possible explanations for this apparent contradiction. First, although the ratio $(\rho + \pi)/\delta$ begins below the critical value $\sqrt{1 - [\tau/(1-k)]}$ in row 1, it ends up above this critical value in row 2. Hence, in raising the rate of inflation from zero to 5 percent, $d\delta/d\pi$ changes sign, and it may be that the negative values of $d\delta/d\pi$ dominate the positive values over this range. Second, Auerbach's table 1 indicates a drop in the real discount rate ρ when going from $\pi = 0$ to $\pi = 5$ percent. The results derived in this paper have assumed that ρ is constant. However, as shown below, a decrease in ρ will tend to cause a decrease in δ, consistent with the results in Auerbach's table 1.

To calculate the effect of ρ on the depreciation rate δ, simply replace λ_π by λ_ρ and λ'_π by λ'_ρ in (11). Differentiating λ with respect to ρ, we obtain

$$\lambda_\rho = \frac{-(1-\tau)A(\delta)\phi}{(\rho+\delta)^2} - \frac{\tau\delta}{(\rho+\pi+\delta)^2} < 0. \tag{22}$$

Note that λ_ρ is negative since an increase in the real discount rate reduces the present value of future rentals (including depreciation deductions) to capital.

Differentiating λ' with respect to ρ, we obtain

$$\lambda'_\rho = \frac{-(1-\tau)\phi}{(\rho+\delta)^2}\left[A'(\delta) - \frac{A(\delta)}{\rho+\delta}\right] + \frac{(1-\tau)\phi A(\delta)}{(\rho+\delta)^3} + \tau\frac{\delta-\rho-\pi}{(\rho+\pi+\delta)^3}. \tag{23}$$

Evaluating this derivative where $\lambda' = 0$, we substitute $\tau(\rho+\pi)/(\rho+\pi+\delta)^2$ for $[-(1-\tau)\phi/(\rho+\delta)]\{A'(\delta) - [A(\delta)/(\rho+\delta)]\}$ to obtain

$$\lambda'_\rho = \frac{(1-\tau)A(\delta)\phi}{(\rho+\delta)^3} + \frac{\tau}{(\rho+\pi+\delta)^3}\left(\delta + \frac{\rho+\pi}{\rho+\delta}\pi\right), \tag{24}$$

which is positive for $\pi \geq 0$. To calculate B_ρ, substitute λ_ρ and λ'_ρ into $B_\rho = -\lambda'_\phi\lambda_\rho + \lambda_\phi\lambda'_\rho$ and use the substitution mentioned in footnote 15 to obtain

$$B_\rho = \frac{(1-\tau)A(\delta)}{(\rho+\delta)(\rho+\pi+\delta)^2}\left\{(q-\tau z)\left[\left(1+\frac{\pi}{\rho+\delta}\right)^2 - 1\right] + \frac{q(q-\tau)}{q-\tau z}\right\}. \tag{25}$$

Under the reasonable assumption that $\tau < 1 - k$, it is the case that $q - \tau > 0$ and, hence, $B_\rho > 0$. Therefore, since $\lambda'_\rho > 0$ and $B_\rho > 0$, we

determine from the appropriately modified version of (11) that $d\delta/d\rho > 0$. Therefore, as mentioned above, a decrease in ρ tends to decrease the optimal value of δ.

We have shown that the direction of the effect of inflation on durability depends on relative magnitudes of the nominal discount rate and the depreciation rate. We now examine the effect of inflation on the steady-state flow of capital services by examining the effect on ϕ, the marginal product of capital services. Using Cramer's rule, we obtain

$$\frac{d\phi}{d\pi} = \frac{1}{\Delta}\left\{(1-\tau)c''g'\frac{I/p}{\delta}\left[A'(\delta) - \frac{A(\delta)}{\delta}\right]\lambda'_\pi - g'\frac{A(\delta)}{\delta}\lambda'_\delta\lambda_\pi\right\}. \quad (26)$$

Note that the coefficient of λ'_π is positive and that the coefficient of λ_π is negative. Furthermore, recall that λ_π is negative. Therefore, $d\phi/d\pi$ is positive if either $c'' = 0$ or λ'_π is positive. Since ϕ is a decreasing function of the flow of capital services, it is clear that, in the absence of adjustment costs, an increase in the rate of inflation reduces the flow of capital services, even though the durability of capital may increase or decrease.[16]

IV. Concluding Remarks

By replacing Auerbach's proportional consumption function with consumption behavior determined by intertemporally optimizing consumers, we can derive strong results about the direction of the effect of inflation on durability without appealing to a simulation analysis. If, in the absence of installation costs, $(\rho + \pi)/\delta$ (the ratio of the nominal discount rate to the rate of depreciation) exceeds a certain critical value, then $d\delta/d\pi < 0$, so that an increase in the rate of inflation leads to more durable capital; if this ratio is less than the critical value, then $d\delta/d\pi > 0$. As explained in Section III, these results are broadly consistent with Auerbach's simulation results in which $d\delta/d\pi < 0$ since Auerbach's simulations are confined to the region in which $(\rho + \pi)/\delta$ exceeds the critical value. However, for less durable capital, the ratio $(\rho + \pi)/\delta$ may fall short of its critical value, and the model presented here indicates that $d\delta/d\pi$ would be positive. Regardless of whether an increase in the rate of inflation increases or decreases durability, we find that, in the absence of installation costs, an increase in the rate of inflation reduces the steady-state flow of capital services.

[16] For the case in which $c'' = \infty$, sign $d\phi/d\pi$ = sign λ'_π = sign $d\delta/d\pi$. In this case, the steady-state rate of investment is invariant to the rate of inflation. If an increase in π increases (decreases) δ, then the steady-state flow of capital services $[A(\delta)/\delta](I/p)$ is reduced (increased), thereby increasing (decreasing) ϕ

References

Abel, Andrew B., and Blanchard, Olivier J. "An Intertemporal Equilibrium Model of Saving and Investment." Report 8006, Univ. Chicago, Center Math. Studies Bus. and Econ., February 1980. Forthcoming in *Econometrica*.

Auerbach, Alan J. "Inflation and the Choice of Asset Life." *J.P.E.* 87, no. 3 (June 1979): 621–38.

Eisner, Robert, and Strotz, Robert H. "Determinants of Business Investment." In *Impacts of Monetary Policy*, compiled by the Commission on Money and Credit. Englewood Cliffs, N.J.: Prentice-Hall, 1963.

Gould, John P. "Adjustment Costs in the Theory of Investment of the Firm." *Rev. Econ. Studies* 35 (January 1968): 47–55.

Lucas, Robert E., Jr. "Optimal Investment Policy and the Flexible Accelerator." *International Econ. Rev.* 8 (February 1967): 78–85.

Treadway, Arthur B. "On Rational Entrepreneurial Behavior and the Demand for Investment." *Rev. Econ. Studies* 36 (April 1969): 227–39.

Uncertain Lifetime, Consumption, and Dissaving in Retirement

James B. Davies

University of Western Ontario

This paper asks whether the continued accumulation, or mild dissaving, observed among the retired can be explained by uncertain lifetime. In the absence of annuities, after an initial period influenced by borrowing constraints, under constant relative risk aversion, uncertain lifetime depresses consumption by a proportion increasing with age if the elasticity of intertemporal substitution in consumption is "small." Illustrative computations, based on actual income and survival data, show that plausible elasticities are sufficiently small to give this effect. The reduction in consumption is large enough to explain much of the lack of decumulation by the elderly.

I. Introduction

It has frequently been observed that the elderly either continue to save in retirement or decumulate much more slowly than would be predicted by the crude life-cycle model without bequest motive under certainty. Most recently, Mirer (1979) has found that, controlling for education, in cross section the mean net worth of American couples continues to rise in retirement.[1] When secular growth is taken into

I would like to thank Glenn MacDonald, Nigel Tomes, Geoffrey Carliner, Chris Robinson, and John Whalley, as well as the participants in the Labor Workshop at the University of Western Ontario and a seminar at McMaster University, for helpful comments. The responsibility for any errors or omissions is my own.

[1] Shorrocks (1975) points out that the slope of the age-wealth profile is biased upward by the lower mortality of the wealthy. Mirer (1979) argues that controlling for education largely corrects for this. While it is not clear that explicit corrections based on estimates of differential mortality by social class provide more reliable results (Shorrocks 1975, p. 163), it should be noted that Shorrocks obtained fairly rapid decumulation in retirement for the United Kingdom by using this alternative.

Journal of Political Economy, 1981, vol. 89, no. 3]
© 1981 by The University of Chicago. 0022-3808/81/8903-0006$01.50

account, this implies an increase, a fortiori, for a cohort. Almost as startling are the positive, or only slightly negative, rates of accumulation found in surveys of saving.[2] However, few, even among the old, *say* they are saving for bequest.[3] Recent work suggests that this may be because human as well as nonhuman intergenerational transfers are possible, and many families are at a corner where only the former are desired (see Drazen 1978; Becker and Tomes 1979; Adams, forthcoming). Thus a bequest-motive explanation of the slow dissaving of the elderly is not as attractive as it might seem.

This paper shows that the observed slow decumulation of the elderly may to a large extent be explained by uncertain lifetime. It does so by developing a model in which the effect of this kind of uncertainty on consumption by persons without a bequest motive can be assessed at different ages. Constant relative risk aversion is assumed for simplicity, but an appendix (obtainable from the author) shows that the results also hold under constant absolute risk aversion. Nonpension annuities are ignored on the argument that a no-bequest life-cycle model is consistent with their observed unimportance in household portfolios only if available annuities are not sufficiently attractive to dominate regular saving instruments for most consumers. With the most plausible taste parameters and rates of return, in this model in the absence of pensions uncertain lifetime will not only depress consumption at all ages but will also have an increasingly severe proportional impact beyond middle age. Illustrative computations, based on actual earnings and survival data, show that this outcome is also likely with realistic pensions and that the effects are sizable. On conservative assumptions, uncertain lifetime may more than halve the mean rate of decumulation among the retired.

The paper is organized as follows. Section II reviews previous work

[2] In three major studies, the following ratios of mean saving to mean net worth for those over 65 were found:

Source	Country	Year	Rate (%)
Lydall (1955, pp. 143, 147)	U.K.	1952	−1.3
Projector and Weiss (1966, p. 110) and Projector (1968, p. 215)	U.S.	1963	+.2
Statistics Canada (1973*a*, p. 84, 1973*b*, p. 140)	Canada	1970	+.8

[3] Only 4 percent of the respondents to the 1962 *Survey of Financial Characteristics* in the United States cited "providing an estate" as a saving objective (Projector and Weiss 1966, table A30). The 1964 Brookings Survey of affluent families (incomes above $10,000) found only 23 percent (of all ages) who were saving to make a bequest (Barlow, Brazer, and Morgan 1966, p. 198).

on the impact of uncertain lifetime on consumption. Section III presents the model, Section IV the illustrative computations.

II. Previous Work

Yaari (1965) studied the impact of uncertain lifetime on the planned rate of change of consumption, examining four cases defined by the presence or absence of a bequest motive and actuarially fair insurance. Except in the cases with a bequest motive, insurance takes the form of either life-insured net borrowing or annuities. Since under actuarial fairness the interest rate on the latter exceeds that on "regular" savings, consumers without a bequest motive hold their entire wealth in the form of annuities.[4]

Omitting a bequest motive, Yaari assumed that a consumer aged t, with probability $P(\tau \mid t)$ of surviving to age τ and a maximum lifetime of T, would have the expected utility integral:

$$V = \int_t^T P(\tau \mid t)\alpha(\tau)U[C(\tau)]d\tau \tag{1}$$

where $C(\tau)$ is consumption at τ, $\alpha(\tau)$ is a subjective discount factor, and the rate of time preference is $1 - P(\tau \mid t)\alpha(t)$.

Maximizing (1) subject to the appropriate constraints yields the following results. In the absence of insurance, since the rate of time preference is higher than under certainty $[P(\tau \mid t) < 1]$, consumption tends to grow more slowly throughout the life cycle. However, actuarially fair insurance makes the price of future consumption $P(\tau \mid t)$ $\exp[-r(\tau - t)]$, where r is the (constant) rate of interest, that is just low enough to offset completely the higher time preference. The result is that, with insurance, consumption grows at about the same rate under uncertainty as under certainty.[5]

It is important to note that Yaari's conclusions concerned only the *shape* of the lifetime consumption path and not its level. Levhari and Mirman (1977, referred to as LM below) have extended Yaari's work by analyzing levels in the absence of insurance. They assume constant rates of subjective discount and relative risk aversion, giving:

[4] The rate of return on an annuity of a year's duration would equal the rate on regular savings plus the probability of death during that year (see Yaari 1965, p. 144).

[5] The actual results are that $\dot{C}(t) = -[r + \hat{\alpha}(t)](U'/U'')$ under both certainty and uncertainty with insurance, while $\dot{C}(t) = -[r + \hat{\alpha}(t) - \hat{P}(t \mid t)](U'/U'')$ under uncertainty without insurance. (A dot denotes a time derivative and a hat a proportional rate of change with time; U' and U'' are first and second derivatives of U with respect to $C[t]$.) Since $C(t)$ will differ in the three cases, and U'/U'' is in general not constant, we can say only, loosely, that under uncertainty consumption *tends* to grow at the same rate as under certainty with insurance but more slowly without.

$$V = \int_t^T P(\tau \mid t) \exp[-\rho(\tau - t)] \frac{C(\tau)^{1-\gamma}}{1-\gamma} d\tau. \tag{2}$$

Here $-\gamma$ is the elasticity of marginal utility with respect to consumption, and γ is the index of relative risk aversion. The elasticity of intertemporal substitution in consumption, σ, is equal to $1/\gamma$.

Maximizing (2) when resources are composed exclusively of nonhuman wealth, LM find that the impact of uncertain lifetime is a priori indeterminate. In the special case where $r = \rho = 0$, initial consumption will be higher than under certainty if $\gamma < 1$ and lower if $\gamma > 1$. When r and ρ are nonzero, there are a number of distinct cases. It is argued in Section IV of this paper that the most "realistic" is where $\gamma > 1$. In this case, unless r or ρ (or both) is very high, increased uncertainty always leads to a decline in initial consumption.[6] As shown below, this conclusion does not extend to the case where resources are largely in "human" form.

Barro and Friedman (1977) have drawn attention to the fact that, under (2), with actuarially fair insurance, consumption grows at the same rate under uncertain lifetime as under certainty. They have also pointed out that if earnings did not change with age, and r and ρ were zero, consumption would be constant throughout life *at the same level* under certainty and uncertainty. This special result appears to be the only available analysis of the impact of uncertain lifetime on the level of consumption, in the presence of insurance.[7]

These previous contributions might be criticized for focusing either on a world without insurance and annuities or on one in which they are actuarially fair. Both situations are clearly "unrealistic." Must we, then, model insurance with an equilibrium load to make further progress? In the present context, the answer is no. Aside from considerations of analytical difficulty, the exercise is not necessary if we wish to discover whether the life-cycle model without a bequest motive can explain the slow dissaving of the old. Nonpension annuities are in fact of little importance for savers of any age.[8] There are two possible

[6] Levhari and Mirman 1977, p. 275. Note that LM did not regard this case as especially significant. In fact, in their abstract they state "the major result is that if the utility function is Cobb-Douglas and the rate of return is not too large relative to the amount of future discounting then lifetime uncertainty will always increase consumption," which refers to the case where $\gamma = 1$.

[7] An appendix to this paper which analyzes the case with actuarially fair insurance is available from the author. The illustrative computations show that, although, theoretically, a positive or negative impact of uncertain lifetime on consumption is possible, a slight negative impact is likely at all ages. There is no tendency, however, for this to be more severe for the elderly.

[8] In the 1962 *Survey of Financial Characteristics* only 1 percent of households reported any private annuities, and on average these composed less than 0.1 percent of the household portfolio. Surprisingly, the figures are not any higher for those aged 65+ (Projector and Weiss 1966, tables A8 and A31).

explanations. Either there is a widespread strong desire to make bequests, or insurance markets fail to provide annuities sufficiently attractive to outweigh the greater transactions costs and inconvenience of saving in this form.[9] Thus the only case in which the life-cycle model without a bequest motive *could* explain the low consumption of the old is that where regular saving instruments generally dominate annuities. Accordingly, for analytical simplicity it is assumed here that regular savings always dominate annuities.

III. The Impact of Uncertain Lifetime at Different Ages

The object of this paper is to determine the impact of uncertain lifetime on consumption at different ages. One might suppose the appropriate procedure would be to compare complete lifetime plans under certainty and uncertainty. Such a comparison would not, however, tell us the impact of uncertainty on an individual of arbitrary age with given current wealth and other characteristics. It would indicate only differences in projected consumption at various ages for individuals with given wealth at the *beginning* of their lifetimes.[10]

The impact of uncertain lifetime on consumption by an individual of age t with wealth, $W(t)$, and a noninvestment income stream, $E(\tau)$, $t \leq \tau \leq T$, can be measured by the difference between what he would consume if his lifetime were certain, $C^c(t)$, and if it were uncertain, $C^u(t)$. (The argument t will be omitted in the text below when no ambiguity arises.) In making this comparison we must vary only the degree of uncertainty. In particular, the expected lifetime,

$$\mu(t) = \int_t^T P(\tau \mid t)d\tau + t, \tag{3}$$

must be held constant.

Consumption under Certainty

Under certainty a consumer of arbitrary age, t, would maximize (2) subject to the constraint that discounted consumption up to $\mu(t)$, $\bar{K}^c(t)$, could not exceed total current resources, $R^c(t)$. Resources would be composed of current wealth, $W(t)$, and the discounted value of nonin-

[9] There is a significant problem of adverse selection since survival chances clearly differ widely but may be quite difficult for an insurance company to ascertain

[10] As mentioned previously, Yaari (1965) showed that under uncertainty consumption tends to grow less quickly than under certainty. Thus, irrespective of the difference in initial consumption (say, at age 20), by retirement, consumption under a complete lifetime plan will almost certainly be lower under uncertainty. This has no bearing, however, on whether an individual aged, say, 65, of given wealth, consumes more or less than he would under certainty.

vestment income, $E(\tau)$, $t \leq \tau \leq \mu(t)$, denoted $H^c(t)$. Adding the terminal condition $W[\mu(t)] = 0$, we have the constraint:

$$W(t) + H^c(t) - \int_t^{\mu(t)} C^c(\tau) \exp[-r(\tau - t)]d\tau = 0, \tag{4}$$

or

$$R^c(t) - \bar{K}^c(t) = 0.$$

This is a standard problem in optimal control. Consumption will grow according to

$$\hat{C}^c(\tau) = \frac{r - \rho}{\gamma} = g, \tag{5}$$

where the hat denotes a proportional rate of change with time (as it will throughout). Equation (5) has the obvious solution:

$$C^c(\tau) = \exp[g(\tau - t)]C^c(t). \tag{6}$$

Substituting (6) into (4) we obtain:[11]

$$C^c(t) = \left\{\int_t^{\mu(t)} \exp[(g - r)(\tau - t)]d\tau\right\}^{-1} R^c(t) = PC^c(t)R^c(t). \tag{7}$$

The term PC^c, the average "propensity to consume" out of resources, is just $PC^c(t) = [\bar{K}^c(t)/C^c(t)]^{-1} = K^c(t)^{-1}$, that is, the inverse of discounted lifetime consumption normalized by initial consumption. It is helpful to view PC^c in this light since it is usually easy to see how a change in uncertainty affects "normalized" lifetime consumption. The latter simply reflects the shape of the consumption path and the strength of discounting.

Consumption under Uncertainty

Under uncertain lifetime, the consumer maximizes (2) subject to the constraint that wealth may never become negative. (To lend without security to a person whose lifetime is uncertain would be to provide a form of insurance.) As Yaari pointed out, under this constraint the optimal plan consists of a series of intervals where the constraint is effective and others where it is not. I refer to these as "blocked" and "free," respectively. In the LM analysis, where all resources are in the form of an initial stock of nonhuman assets, the entire lifetime is a free interval. Here matters are more complicated.

Over blocked intervals we have, of course, $C^u = E$. In addition, it is clear that the consumption plan is continuous at the beginning, t^*,

[11] Eq. (7) is not simplified, in order to emphasize its similarity to the corresponding expressions under uncertainty, eqq. (10) and (13).

and end, $t^* + m$, of any free interval, with $C^u(t^*) = E(t^*)$ and $C^u(t^* + m) = E(t^* + m)$.[12] Over the course of a free interval, consumption changes according to

$$\hat{C}^u(\tau) = \frac{1}{\gamma}[r - \rho + \hat{P}(\tau)] \tag{8}$$

(Yaari 1965, p. 143).[13] That is, consumption grows less quickly than under certainty because expected utility from future consumption declines without any corresponding drop in price. Solving (8), we obtain:

$$C^u(\tau) = P(\tau \mid t)^{1/\gamma} \exp[g(\tau - t)]C^u(t), t^* \leq t \leq \tau \leq t^* + m. \tag{9}$$

Recalling that wealth must be exhausted over the interval, we have:

$$C^u(t^*) = \frac{\int_{t^*}^{t^*+m} E(t) \exp[-r(t - t^*)]dt}{\int_{t^*}^{t^*+m} P(t \mid t^*)^{1/\gamma} \exp[(g - r)(t - t^*)]dt}. \tag{10}$$

Equations (9) and (10) completely specify the free interval when $t^* + m = T$. In the case where $t^* + m < T$, we have a further relationship. Since $C^u(t^*) = E(t^*)$ and $C^u(t^* + m) = E(t^* + m)$, given (9), $E(t^*)$ and $E(t^* + m)$ are related as follows:

$$E(t^*) = \frac{E(t^* + m)}{P[(t^* + m) \mid t^*]^{1/\gamma} \exp(gm)}, m < T - t^*. \tag{11}$$

An indication of the range of plans that may be obtained with realistic E and P paths and plausible values of r, ρ, and γ is provided by figure 1. The plans shown are the result of the illustrative computations of the next section. Note that in all cases life begins with a blocked interval and there is just one free interval.[14] Part A indicates that with $E = 0$ beyond age 65 (no pensions), the free interval would extend until T. That is, wealth would not be run to zero as long as there was a nonzero probability of surviving any longer. In part B, the consequences of realistic pensions are shown. With a high γ, the consumption path is relatively flat (see [9]) and a free interval lasting

[12] If $C^u(t^*) < E(t^*)$, expected utility could be increased by raising consumption at t^* and reducing it just before. A similar argument holds for $t^* + m$.

[13] Note that, since $P(\tau \mid \tau) = 1$, $\hat{P}(\tau) = \hat{P}(\tau \mid \tau)$.

[14] In general, the consumption plan will not always have these features. The initial blocked interval is caused here by E being low at the outset. A second free interval does not succeed the first because, beyond $t^* + m$, the desired growth rate of consumption is declining because of the rapid fall of P, whereas pensions (if present) fall at a constant rate of 2.25 percent per year. (See the next section for an explanation of this rate.) An increase in the ratio of the growth rates of desired consumption and E would be required for a second free interval to emerge.

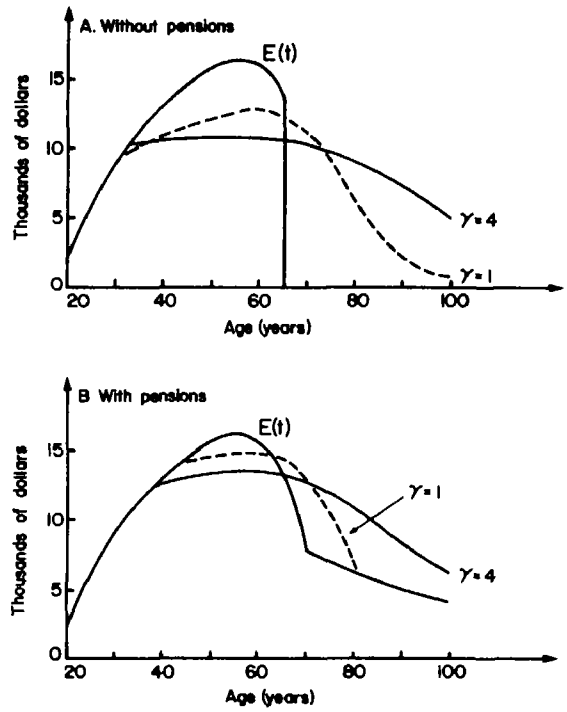


FIG. 1.—Selected optimal consumption paths, $r = .03$, $\rho = .015$

until T is obtained, whereas with $\gamma = 1$ the desired growth rate of consumption declines quickly and a second blocked interval occurs.

Behavior of C^u/C^c *if Free Interval Lasts until* T

How does C^u compare with C^c at different ages? Consider, first, situations where the free interval lasts until T. At any point in the free interval, discounted consumption up to T cannot exceed total "resources," R^u, that is, the sum of current wealth and the present value of the E stream up to T, H^u. Adding the terminal condition $W(T) = 0$, we have:

$$W(t) + H^u(t) - \int_t^T C^u(\tau) \exp[-r(\tau - t)]dt = 0, \tag{12}$$

or $R^u(t) - \bar{K}^u(t) = 0$. Substituting (9) into (12), we obtain:

$$C^u(t) = \left\{\int_t^T P(\tau \mid t)^{1/\gamma} \exp[(g - r)(\tau - t)]d\tau\right\}^{-1} R^u(t) = PC^u(t)R^u(t). \tag{13}$$

Again, $PC^u(t) = K^u(t)^{-1}$.

TABLE 1

ANALYSIS OF PC^u/PC^c

g and r	γ	PC^u/PC^c
$g > r$	<1	>1 if γ sufficiently below 1, otherwise <1
	$=1$	<1
	>1	<1
$g = r$	<1	>1
	$=1$	$=1$
	>1	<1
$g < r$	<1	>1
	$=1$	>1
	>1	<1 if γ sufficiently large, otherwise >1

The behavior of C^u/C^c is determined by that of R^u/R^c and PC^u/PC^c. When the free interval lasts until T, in the presence of pensions we have $R^u/R^c > 1$. However, as will be shown, we may have $PC^u/PC^c < 1$, so that $C^u/C^c < 1$ is a possible outcome.

From equations (7) and (13), we see that PC^u/PC^c depends on the relative size of g and r and the value of γ. In Section IV it is argued that γ is typically considerably greater than unity. It is observed that apparently non-credit-constrained households consume at an increasing rate. As long as utility is additively separable over time, this is only consistent with an interest rate greater than the subjective discount rate.[15] Thus realistic restrictions include $\gamma > 1$ and $\rho < r$. Further, if γ is typically considerably above unity, unless ρ is far below zero, which seems unrealistic, we have $g < r$.

Levhari and Mirman have provided a full analysis of the relative values of PC^u and PC^c. Their conclusions are set out in table 1. When $g = r$ and $\gamma = 1$, recalling (3), we find that (7) and (13) imply that $PC^u = PC^c$. Lowering γ below 1 raises $1/\gamma$ and increases PC^u/PC^c, while raising it does the opposite. With $g > r$ and $\gamma = 1$, $PC^u/PC^c < 1$ since greater relative weight is placed on the more remote future in the denominator of PC^u than in that of PC^c. Raising γ simply reinforces this. On the other hand, although reducing γ below 1 tends to increase PC^u/PC^c, the ratio will not rise above unity until γ is sufficiently small, so that a simple result is not obtained.

Now, in the apparently realistic situation where $g < r$, we have $PC^u/PC^c > 1$ with $\gamma \leq 1$, again because more weight is placed on the more remote future in the denominator of PC^u. However, raising γ

[15] In Yaari's framework, which does not require constant risk aversion or constant subjective discounting, consumption will not rise unless the rate of return exceeds the subjective rate of discount; see Yaari (1965), pp 143, 147.

above unity reduces $1/\gamma$ and tends to make PC^u fall. Clearly, with sufficiently large γ we obtain $PC^u/PC^c < 1$. In the next section we find that with actual earnings and survival data, γ does not have to be far above unity for this result to occur.

In order to study changes in PC^u/PC^c with age, one may compare the proportional rates of change:[16]

$$\widehat{PC^c}(t) = PC^c(t)(1 - \exp\{(g - r)[\mu(t) - t]\}\dot{\mu}(t)) + (g - r), \quad (14)$$

and

$$\widehat{PC^u}(t) = PC^u(t) + \frac{\dot{P}(t)}{\gamma} + (g - r). \quad (15)$$

To interpret these relationships, it is helpful to note that $\widehat{PC^i} = -\hat{K}^i$. Consider first a situation where the lifetime was actually fixed and $g = r$. In this case, normalized lifetime consumption, K^i, would simply equal the remaining life expectancy, $\mu(t) - t$. Each year K^i would fall by a percent equal to its own inverse. Hence $PC^i = (K^i)^{-1}$ would *rise* by this same percent, that is, PC^i itself.

When allowing $g \neq r$, say $g < r$, we give K^i some tendency to rise as t increases since weaker discounting (compounding, if $g > r$) is applied. This gives rise to the $(g - r)$ in both (14) and (15). Finally, in fact lifetime is not fixed. Under certainty, the extension in possible lifetime increases normalized lifetime consumption, reducing PC^c, while under uncertainty, the increased probability of survival to any future date also raises K^i and reduces PC^i.

Equations (14) and (15) show that, in the main, whether PC^u/PC^c tends to rise or fall depends on whether it starts out greater or less than unity. When $PC^u/PC^c < 1$, a decline *may* not necessarily occur. However, since $\dot{\mu}(t)$ and $\dot{P}(t)$ are both typically quite small, in our "most realistic" case, of $g < r$ and γ well above unity, it is likely that PC^u/PC^c will be low enough to ensure its continuous decline. Whether this translates into a decline in C^u/C^c depends entirely on the behavior of R^u/R^c.

In the absence of pension income, since most people retire before the age at which death was expected, even from the viewpoint of the start of adult life, R^u is typically the same as R^c. Hence everything said above about PC^u/PC^c applies to C^u/C^c. That is, with the most plausible parameters $C^u/C^c < 1$ and declines monotonically with age over the free interval. In addition, it is clear that $C^u/C^c < 1$ during the initial blocked interval (C^u will be less than the "free" value given by eq. [13]), although the ratio would likely typically *increase* with age prior to t^* since E rises quickly for the young. (Recall $C^u = E$ in the blocked

[16] In deriving (15) let $\pi(\tau)$ = the unconditional probability of surviving to age τ. Then $P(\tau|t) = \pi(\tau)/\pi(t)$, and $\partial P(\tau|t)/\partial t = -\pi(\tau)\pi(t)^{-2}\dot{\pi}(t)$, or $-P(\tau|t)P(t)$.

interval.) These conclusions are particularly interesting since many of the studies indicating slow decumulation by the old were conducted when pensions were less sizable and widespread than they are today.

When pension income is present, we of course have $R^u/R^c > 1$; and even when the free interval continues to last until T, the above conclusions will hold only with a higher value of γ. In the next section we find that the required level of γ is not, however, implausible. Here we should note that, although $R^u/R^c > 1$, it has some tendency to fall with age. The R^t has two components, W and H^t, the former of which we hold constant in the comparison of certainty and uncertainty. Since most often W is rising and H^t is falling, R^u and R^c have a basic tendency to become more equal. In the next section we find that in practice this dominates, and R^u/R^c typically falls with age. It should be noted, however, that the behavior of H^u/H^c is in opposition. We have

$$\hat{H}^u(t) = r - \frac{E(t)}{H^u(t)}, \tag{16}$$

and

$$\hat{H}^c(t) = r + \frac{E[\mu(t)] \exp\{-r[\mu(t) - t]\}\dot{\mu}(t) - E(t)}{H^c(t)}. \tag{17}$$

Since $E[\mu(t)] \exp\{-r[\mu(t) - t]\}\dot{\mu}(t)$ is usually small relative to $E(t)$ and $H^u > H^c$, this means that H^u tends to fall more slowly than H^c.

Behavior of C^u/C^c *if Free Interval Ends before* T

As we saw in figure 1, in the presence of pensions the free interval may terminate before T. In this case, it is likely that $C^u/C^c > 1$ throughout the free, as well as the following blocked, intervals.

In the computations of the next section, in cases where the free interval ends before T, we usually find that $\mu(t^*) < t^* + m$. Hence $\mu(t)$ must rise and become equal to $t^* + m$ at some point in the interval. At this age C^u exceeds C^c. In general, we have:

$$C^u(t) = \frac{W(t) + \int_t^{t^*+m} E(\tau) \exp[-r(\tau - t)]d\tau}{\int_t^{t^*+m} P(\tau \mid t)^{1/\gamma} \exp[(g - r)(\tau - t)]d\tau}, \quad t^* \leq t < t^* + m, \tag{18}$$

and

$$C^c(t) = \frac{W(t) + \int_t^{\mu(t)} E(\tau) \exp[-r(\tau - t)]d\tau}{\int_t^{\mu(t)} \exp[(g - r)(\tau - t)]d\tau}, \quad 0 \leq t \leq T. \tag{19}$$

For $\mu(t) = t^* + m$ relevant resources are the same, but the denominator is smaller under uncertainty, and $C^u > C^c$.

What can we say about C^u/C^c at other points? Before the age where $\mu(t) = t^* + m$, the longer horizon under uncertainty makes the denominator of (18) larger relative to that of (19). However, relevant human resources in the numerator of (18) are greater than those in (19). Similarly, above the point where $\mu(t) = t^* + m$, the denominator of (18) falls relative to that of (19) but the numerator does the same. In the illustrative computations, the net result is often that $C^u > C^c$ throughout the free interval.

Finally, after the free interval we are likely to have $C^u/C^c > 1$ again. To see this, consider a case where $t^* + m$ occurs in retirement and the pension changes at a constant percentage rate. As long as g exceeds this rate, which appears likely, positive saving will be desired under certainty, and $C^c < E = C^u$.

IV. Quantitative Impact

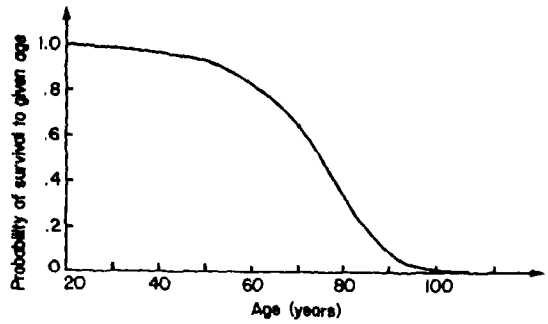
This section provides a quantitative study of the likely impact of uncertainty on consumption at different ages. It does this by computing hypothetical consumption plans under certainty and uncertainty at different ages, using actual data on noninvestment income and survival probabilities and a range of values of r, ρ, and γ. The first step is to compute the complete lifetime consumption plan under uncertainty. This generates a time path for $W(t)$. Next, at intervals of 1 year I ask how much a person would consume, given the $W(t)$ computed in the first step and a *certain* lifetime equal to life expectancy at that age.

Figures 2 and 3 plot the data used. An inverse logistic curve fits the P path in figure 2 well and allows the use of continuous time in the computations.[17] Similarly, the lifetime path of mean noninvestment income in figure 3 is approximated from ages 20 to 70 by a fourth-order polynomial.[18] Beyond age 70, it is assumed that E declines at the observed trend rate of 4.25 percent per year. It is also assumed that secular growth shifts the entire cross-section profile upward at a rate of 2 percent per year. This explains why the E path for an *individual* shown in figure 1 differs from the cross-section profile of figure 3.

The computations were carried out with alternative values of r, but I focus mainly on the case where $r = .03$, suggested by popular estimates of mean household rates of return (see, e.g., Boskin 1978, p. S19; Feldstein, Green, and Sheshinski 1978, p. S64). The subjective

[17] The curve $P(t) = [(1 + a) \exp(-bt)]/[1 + a \exp(ct)]$ was fitted by least squares, using numerical methods ($R^2 = .999$). Values of the parameters are $a = 1.00031$, $b = -0.002114$, and $c = 0.14651$.

[18] The curve, $E(t) = -36{,}999.4 + 3520.22t - 101.878t^2 + 1.34816t^3 - 0.00706233t^4$, was fitted by ordinary least squares ($R^2 = .999$).



Source Statistics Canada (1974, pp 14-15)

Source. Revenue Canada (1973, pp 50-59)

FIGS. 2–3.—Fig. 2, Male survival probabilities (Canada 1971). Fig. 3, Average male noninvestment income (Canada 1971).

rate of discount is limited to the arguably "realistic" range $0 \leq \rho \leq r$. (It was suggested in the previous section that $\rho > r$ is implausible, while allowing $\rho < 0$ simply reinforces the results obtained below.) The parameter γ is allowed to vary from 0.5 to 5.0, reflecting both a desire to cover the alternatives $\gamma \lesseqgtr 1$ and considerable evidence that γ is typically much above unity.

A variety of evidence on γ is available. Friend and Blume (1975, p. 920) estimated it in the context of risk aversion, using data on household asset holdings, and found that it was likely to be in excess of 2.0. Farber (1978), attempting to infer the preferences of members of the United Mine Workers from the results of collective bargaining in the coal industry, also estimated γ as the index of relative risk aversion, obtaining alternative estimates of 3.0 and 3.7 (p. 935). Further evidence is provided by Ghez and Becker (1975, p. 133) and MaCurdy (1979, p. 34), who, in studies of the allocation of goods and time over the life cycle, estimated upper bounds on σ of .28 and .33, re-

TABLE 2

C^u/C^c WITHOUT PENSIONS

	γ							
	$r = .03$						$r = .015$	$r = .045$
AGE	.5	1	2	3	4	5	4	4
				$\rho = 0$				
20	1.060*	.441*	.306*	.274*	.260*	.252*	.236*	.285*
30	1.264	1.000	.903	.872	.857	.835*	.794	.886*
40	1.305	1.000	875	.833	.811	.798	.748	.863
50	1.375	1.000	.832	.775	.746	.728	.684	.800
60	1.498	1.000	.771	695	.656	.633	.600	.709
70	1.678	1.000	.702	609	.564	.538	.518	.609
80	1.815	1.000	668	571	.525	.499	.490	559
				$\rho = r/2$				
20	.441*	.306*	.260*	.247*	241*	237*	226*	257*
30	1.216	986	.857*	.820*	.802*	792*	.798*	810*
40	1.261	1.028	903	.855	.830	814	760	.885
50	1.342	1.032	862	.798	.765	.744	.695	.823
60	1.481	1.038	.802	.718	.674	.648	.610	.733
70	1.675	1.041	.731	.629	579	550	.526	.630
80	1.823	1.039	.691	.586	.537	.508	.496	576
				$\rho = r$				
20	224*	.224*	224*	.224*	.224*	.224*	.217*	.234*
30	752*	752*	.752*	752*	752*	.752*	.769*	743*
40	1.064*	1.044	927	.876	847	829	.771	.904
50	1.293	1.054	.889	820	782	.759	.705	845
60	1.446	1.067	.830	.739	.692	.662	619	756
70	1.657	1.076	.758	.649	594	.562	.533	.652
80	1.821	1.074	.714	602	.548	517	502	.593

* Nonnegative wealth constraint effective under uncertainty

spectively, implying $\gamma \geqslant 3.6$. Finally, recent studies of the elasticity of aggregate saving with respect to the real rate of interest suggest high values of γ as well. Boskin (1978, p. S4), for example, estimates the uncompensated elasticity to lie in the range 0.3–0.4 in the United States. Including pensions, with $r = .03$ and $\rho = .015$, my model implies uncompensated aggregate saving elasticities of 0.9, 0.5, and 0.3 with $\gamma = 3$, 4, and 5, respectively.[19]

To sum up: The available evidence suggests that γ may typically be above 3; that a value of 4 may be a "best guess" at the true figure; and that even $\gamma = 5$ could not be rejected out of hand as implausible.

Table 2 presents the illustrative computations for the case where E is set to zero beyond age 65, that is, where there are no pensions. A value of γ below 1 gives $C^u/C^c > 1$ and rising with age in the free

[19] Alternative values of r and ρ yield similar results. Using $\gamma = 4$, with $r = .06$ and $\rho = .03$ the elasticity is 0.6, while with $r = .03$ and $\rho = .0$ we get a figure of 0.4, e.g.

TABLE 3

C^u/C^c WITH PENSIONS

	γ							
	$r = .03$						$r = .015$	$r = .045$
AGE	.5	1	2	3	4	5	4	4
				$\rho = 0$				
20	.997*	.415*	288*	.258*	245*	.237*	.217*	.274*
30	1.346	1.089	.909*	.827*	.790*	.769*	.740*	.836*
40	1.397	1.109	.992	.946	.922	.908	.915	.935
50	1.491	1.145	.988	922	.888	.867	.886	.899
60	1 669	1 211	.983	.885	.835	.805	.842	.840
70	1.965	1.337	994	.844	772	.730	.795	.766
80	1.808	1.500	1 065	.843†	748†	.696†	.793†	.726†
				$\rho = r/2$				
20	.415*	288*	.245*	.232*	.227*	.223*	.207*	.247*
30	1.232*	.909*	.790*	.756*	.739*	.730*	.714*	.764*
40	1.330	1.168	1.043*	974	945	.926	929	960
50	1.437	1.178	1.030	959	.917	.891	.904	.933
60	1 618	1.266	1 046	937	.874	.835	.865	.888
70	1.705	1 397	1 090	.917	.822	767	.825	.827
80	1.222*	1 230	1 193	943†	807†	.737†	.830†	.798†
				$\rho = r$				
20	.210*	.210*	.210*	.210*	.210*	.210*	.199*	.224*
30	693*	.693*	.693*	.693*	.693*	.693*	.688*	.701*
40	.935*	.935*	.935*	.935*	935*	.935*	.943	925*
50	1.138*	1 138*	1.056	.991	.944	.914	.921	.964
60	1.398*	1 263	1.089	.985	.912	866	.889	.934
70	1.123*	1.318	1.154	992	875	.807	.857	.895
80	1.088*	1.088*	1.220	1.053	.876†	784†	.871†	.887†

* Nonnegative wealth constraint effective under uncertainty.
† No blocked interval succeeds free interval.

interval, while $\gamma > 1$ yields the opposite, as predicted in the previous section. With $\gamma \geq 3$ dramatic results are obtained. Using the "best guess" figure of $\gamma = 4$ and the combination $r = .03$ and $\rho = .015$, for example, maximum negative impacts on consumption are 23.5 and 46.3 percent for the middle-aged and old, respectively. With the same, quite conservative, parameters, the mean ratio of dissaving to net worth for those aged 65–85 declines from 9.0 to 3.7 percent with the introduction of uncertainty. Note, finally, that the results are robust with respect to changes in r.

Table 3 shows that when noninvestment income in retirement is introduced, the ratio C^u/C^c rises in most cases. However, with $\gamma \geq 3$ and $\rho < r$, the impact of uncertainty remains negative and much

larger for the retired than for the middle-aged. When we focus again on the $r = .03$, $\rho = .015$, $\gamma = 4$ case, we see that, while uncertainty reduces consumption in middle age by up to 8.3 percent, it depresses that in retirement by up to 19.3 percent. Although this is a more modest effect than obtained without pensions, it is still impressive, as shown by the fact that with the same parameter values the average rate of decumulation between ages 65 and 85 falls from 7.0 to 2.9 percent with the introduction of uncertainty.[20] Furthermore, isolation of the $r = .03$, $\rho = .015$, $\gamma = 4$ case is conservative. Small changes in parameter values can make the effect even more sizable. By reducing ρ to 0 or by raising γ to 5, for example, we increase the maximum negative impact on consumption for the elderly to about 25 percent and reduce the mean rates of decumulation under uncertainty for those aged 65–85 to 2.4 and 2.1 percent, respectively. Finally, note that the results are again robust with respect to r.[21]

V. Conclusion

This paper has shown that the life-cycle model without a bequest motive can help to explain one of the most interesting features of actual consumption behavior, the typical low rate of dissaving among the elderly, when uncertain lifetime is taken into account. Theory shows that, with the most plausible parameters, a negative impact of uncertain lifetime which increases in proportional severity with age is likely in the absence of pensions and may also occur when pensions are present. Illustrative computations based on actual income and survival data show that this effect is, in fact, obtained with or without pensions. Furthermore, the magnitude of the impact is sufficient to suggest that uncertain lifetime could provide the major element in a complete explanation of the slow decumulation of the retired.

References

Adams, James D. "Personal Wealth Transfers." *Q.J.E.* (forthcoming).

Barlow, Robin; Brazer, Harvey E.; and Morgan, James N. *Economic Behavior of the Affluent.* Washington: Brookings, 1966.

[20] The contrast in decumulation rates remains so striking because a given proportional difference in consumption implies a *larger* proportional difference in saving when $E > 0$ because the scale of private wealth holding and dissaving is reduced.

[21] White (1978, p. 559) states that she (1976, chap. 5) demonstrated that the no-bequest life-cycle model without insurance cannot explain aggregate personal saving in the United States. Indeed, the simulations of White (1976, p. 115) show consumption increasing as a result of uncertain lifetime in all cases considered. The explanation of the apparent conflict with the present paper is that White examined only values of γ between 0.5 and 1.5. Thus her results correspond to those shown here in the initial columns of table 3.

Barro, Robert J., and Friedman, James W. "On Uncertain Lifetimes." *J.P.E.* 85, no. 4 (August 1977): 843–49.

Becker, Gary S., and Tomes, Nigel. "An Equilibrium Theory of the Distribution of Income and Intergenerational Mobility." *J.P.E.* 87, no. 6 (December 1979): 1153–89.

Boskin, Michael J. "Taxation, Saving, and the Rate of Interest." *J.P.E.* 86, no. 2, pt. 2 (April 1978): S3–S27.

Drazen, Allan. "Government Debt, Human Capital, and Bequests in a Life-Cycle Model." *J.P.E.* 86, no. 3 (June 1978): 505–16.

Farber, Henry S. "Individual Preferences and Union Wage Determination: The Case of the United Mine Workers." *J.P.E.* 86, no. 5 (October 1978): 923–42.

Feldstein, Martin S.; Green, Jerry; and Sheshinski, Eytan. "Inflation and Taxes in a Growing Economy with Debt and Equity Finance." *J.P.E.* 86, no. 2, pt. 2 (April 1978): S53–S70.

Friend, Irwin, and Blume, Marshall E. "The Demand for Risky Assets." *A.E.R.* 65 (December 1975): 900–922.

Ghez, Gilbert R., and Becker, Gary S. *The Allocation of Time and Goods over the Life Cycle.* New York: Columbia Univ. Press (for Nat. Bur. Econ. Res.), 1975.

Katz, Eliakim. "A Note on Uncertain Lifetimes." *J.P.E.* 87, no. 1 (February 1979): 193–96.

Levhari, David, and Mirman, Leonard J. "Savings and Consumption with an Uncertain Horizon." *J.P.E.* 85, no. 2 (April 1977). 265–81.

Lydall, Harold F. "The Life Cycle in Income, Saving, and Asset Ownership." *Econometrica* 23 (April 1955): 131–50.

MaCurdy, Thomas E. "An Empirical Model of Labor Supply in a Life Cycle Setting." Nat. Bur. Econ. Res. Working Paper no. 421, December 1979.

Mirer, Thad W. "The Wealth-Age Relation among the Aged." *A.E.R.* 69 (June 1979): 435–43.

Projector, Dorothy S. *Survey of Changes in Family Finances.* Washington: Board Governors, Federal Reserve Board, 1968.

Projector, Dorothy S., and Weiss, Gertrude S. *Survey of Financial Characteristics of Consumers.* Washington: Board Governors, Federal Reserve Board, 1966.

Revenue Canada. *1973 Taxation Statistics.* Ottawa: Minister Supply and Services, 1973.

Shorrocks, A. F. "The Age-Wealth Relationship: A Cross-Section and Cohort Analysis." *Rev. Econ. and Stats.* 57 (May 1975): 155–63.

Statistics Canada. *Family Expenditure in Canada 1969.* Publication no. 62-535. Ottawa: Ministry Indus., Trade, and Commerce, 1973. (*a*)

———. *Incomes, Assets and Indebtedness of Families in Canada.* Publication no. 13-547. Ottawa: Ministry Indus., Trade, and Commerce, 1973. (*b*)

———. *Life Tables: Canada and Provinces.* Publication no. 84-532. Ottawa. Ministry Indus., Trade, and Commerce, 1974.

White, Betsy Buttrill. "On the Rationality of Observed Saving: A Critique of the Life Cycle Hypothesis." Ph.D. dissertation, Stanford Univ., 1976.

———. "Empirical Tests of the Life Cycle Hypothesis." *A.E.R.* 68 (September 1978): 547–60.

Yaari, Menahem E. "Uncertain Lifetime, Life Insurance, and the Theory of the Consumer." *Rev. Econ. Studies* 32 (April 1965): 137–50.

What Kind of a Science Is Economics? A Review Article on *Causality in Economics* by John R. Hicks

Christopher A. Sims

University of Minnesota

As computational costs have declined, economists have used increasingly complicated statistical models in analyzing business cycles. Most of these models have been based on extensions of the IS-LM framework developed by Hicks, combined with the simultaneous equations statistical framework which was the central subject matter of econometrics. These large econometric models were, and still are, successful in certain important respects. By making possible the systematic use of large amounts of statistical information and the codification of forecasting and projection procedures, they provided good forecasts—as good as or better than averages of "judgmental" forecasters. The models and their services have become commercially valuable. Nonetheless, they seem to many economists to have outgrown their underpinnings in economic and statistical theory.

The economic theory entering the models does not in itself explain the existence of recurrent business cycles. In practice, the fitted models usually explain much of what might be labeled "the character of the business cycle" as a consequence of properties of the "error terms" in the model about which theory has little to say. This is not entirely the fault of the economic theory. The simultaneous equations statistical theory makes heavy demands in this area of application on supposed a priori knowledge. It is therefore perhaps not surprising that the actually available economic theory leaves large gaps to be filled with conventional or ad hoc assumptions. The burden of ad hoc assumptions and implicit statistical procedures carried by the models for these reasons weakens them somewhat as tools for forecasting and

[*Journal of Political Economy*, 1981, vol. 89, no. 3]
© 1981 by The University of Chicago 0022-3808/81/8903-0009$01.50

policy projection. It weakens them still more as scientific instruments when important economic issues are subject to dispute.

Reflecting these sources of dissatisfaction are two current developments in macroeconomic research. On the one hand, some economists are building stochastic theoretical models which have testable probabilistic implications for the behavior of major aggregate time series as stochastic processes. On the other hand, econometricians have been looking at methods for reducing the ad hockery of the standard methodology by formalizing the procedures for controlling the dimensionality of models and for accounting for drift in model structure.

These developments seem to me natural and hopeful. They do, though, point to an increasingly complicated and intense interaction between theory and data, using probabilistic models. What is going on is not quite like what has gone on in any other science, though one can see some analogies. Is what is going on useful? What are the criteria for confirming or disconfirming models of this type? Are the formal and informal standards for reporting results in this area reasonable?

According to its introduction, the book under review grew out of Hicks's meditation on a 1974 conference on the microfoundations of macroeconomics, which he attended and saw as a failure. It also reflects his view that "econometrics is now in some disarray." The book discusses scientific method in general and then in this light the special character of economics, especially macroeconomics, as science. As is evident from the book's length (124 + xii pages), the argument is concise, and it is presented in a conversational tone. The central conclusions, as I see them, are that "as economics pushes on beyond 'statics' it becomes less like science, and more like history" (p. xi) and that "the probability calculus is no excuse for forgetfulness" (p. 122).

To explain completely what Hicks means by these phrases would require more space than I have in this review, but some explication is obviously required. The former idea is that dynamic economic theories must inherently be incomplete, imprecise, and therefore subject to variation over time. One reason for this is that economic cause-effect relations involve a "recognition delay" about which theory has little to say and which may be expected to be variable. For example, when a price change causes a change in an amount purchased, there will be a delay between the actual price change and consumers' hearing of it, as well as a possible further delay before consumers decide to act on the information. It is wrong, then, to expect economic theories to be complete, mechanical, and divorced from reference to specific historical circumstances.

The latter idea is that, particularly in economic time series, probability models are often inappropriate because the data are interde-

pendent in ways which make treating them as "random" patently unjustifiable. Hicks argues that probability judgments in economics, both in microeconomic decision making and in evaluation of competing economic theories, often result in incomplete orderings, with many possibilities lying in a gray zone of noncomparable probabilities. Thus the usual axiomatic arguments for the probability calculus are unjustified in a wide range of economic applications. He cites as an example of this objection to the probability calculus in inference the situation in which a new fact which seems not to fit existing theory emerges, and a new but ad hoc theory is put forward to "explain" it. Because a priori and likelihood considerations point in different directions in this case, he argues that the two theories lie in the gray zone of noncomparability and that the reasonable action to take is to reserve judgment. Similar considerations, Hicks argues, suggest that economic behavioral models of decision making under uncertainty ought not to be formulated taking the usual probability calculus for granted. Probabilistic models and methods of inference are often awkward and misleading tools in application to economics.

Hicks's discussion is valuable for the insights it presents into the special character of economics as a science. Its conclusions, though, seem to me deeply mistaken, in part because the argument too easily resolves some of the problems it raises.

My dissatisfaction with the Hicks book is easier to explain if we compare it with another recent book on a closely related subject—*Chance, Cause, Reason*, by Arthur W. Burks (1977). The Burks book is expansive (almost 700 pages), where Hicks's is concise; it is densely formal in parts, where Hicks's is persistently informal; it very seldom gives the impression of having chosen an easy resolution of a difficulty, where Hicks's does. Burks, for example, considers at some length the objection to the probability calculus given central place by Hicks: that probability judgments do not yield a complete ordering. Burks argues that these objections are correct in principle but that they are in the nature of qualifications to the range of applicability of the calculus—when no decision has to be made, or when the decision to be made is unimportant, then it may not be "economic" (to use Burks's word) to rank all probabilities and utilities. But (this is my conclusion, not Burks's) in important issues of scientific inference or in economic decision making which involves substantial amounts of money and utility, it is both normatively and descriptively reasonable to treat probabilities and utilities as completely ordered. In taking decisions, we implicitly rank them in any case.

I could go on citing Burks in rebuttal to parts of Hicks's argument with which I disagree, but it would be better for the reader of this review to read Hicks and struggle with Burks (I for one cannot "read"

Burks in the ordinary sense of the term) and reach his own conclusions. For despite my sympathy with Burks's style and with many of his arguments, I do not see his book as in any sense a substitute for or a refutation of Hicks. Burks is constructing a model of scientific inference, and it seems that the hard sciences are the focus of his interest. He rightly points to the need for some degree of uniformity in scientific laws over space and time if inference is to be possible and also rightly (in my view) claims that his model of inference applies even if there is some variation of laws in space and time. He rightly claims that there is a sense in which empirical probability statements (e.g., "the probability that this coin will come up heads is .5") can be true, and from this point builds up an intersubjective (as opposed to Savage's personalistic) theory of inference. Economists do often aim at influencing professional consensus as to what causal structure plausibly underlies historical data. If we are to think about how economists do this or ought to do this, an intersubjective theory like Burks's fills an important gap in Savage's subjective theory.

But Hicks is right about some of the central differences between economics and other sciences—temporal instability of our theories is not a marginal qualification to our methods of inference but a central aspect of our difficulties in determining what is true. And treating "empirical truth" as a primitive notion seems too simple to lead to a theory applicable to economics. The problem of the "ad hoc theory," arising from the dearth of "empirically true" probability statements in economics and the near impossibility of generating them by experiment, is central to us. Burks's system, while relevant to economics, fascinating on its own terms, and stimulating as an example, does not confront in detail some of the central problems of inference in economics.

Hicks is not explicit about how his broad conclusions apply to specific types of current work in economics. (Indeed, except for the author's own works, no publications in macroeconomics, probability, or statistics less than 15 years old are cited.) He does in his final paragraph appear to condemn all use of probability methods on cross sections not generated by random sampling:

> Thus it is not at all sensible to take a small number of observations (sometimes no more than a dozen observations) and to use the rules of probability theory to deduce from them a "significant" general law. For we are assuming, if we do so, that the variations from one to another of the observations are random, so that if we had a larger sample (as we do not) they would by some averaging tend to disappear. But what nonsense this is when the observations are derived, as

> not infrequently happens, from different countries, or localities, or industries—entities about which we may well have relevant information, but which we have deliberately decided, by our procedure, to ignore. [Pp. 121–22]

I think Hicks is wrong here. Any use of experience to reach better decisions must involve "deliberately ignoring" information, whether the experience is the historical record or a sequence of repeated experiments. In physical experiments there is always observational error, which is treated as random. A good experimenter may have many bits of information he judges relevant to explaining the apparent random error in a sequence of experiments. In writing up the experiments, though, he does not report which research assistant was on duty at each meter reading, together with an analysis of each assistant's bias and mean square deviation as a meter reader. In physical sciences there is substantial consensus as to what is to be treated as random.

In social sciences, with history the most extreme instance, there is much less consensus. It is nonetheless necessary to ignore some information, to treat it as unsystematic or random, if the past is to be useful. Because there is inevitably room for dispute about what is systematic and what is not in analyzing history, historians do not dogmatically suppress discussion of information which some might consider "random" or unsystematic. Nonetheless, when a decision has to be made, if history is to be relevant, some choice must be made as to what is random. A political decision maker might well need to judge how likely it is that a Marxist revolution in his country would be followed by political democracy. He might think that the historical record concerning previous Marxist revolutions is relevant to this question. However, it is obvious that every previous revolution has special characteristics which are well known and which are potentially relevant to explaining whether the postrevolutionary government emerged as politically democratic. In applying the historical record to his own problem, the decision maker will have to treat at least some of these special characteristics as random.

The same type of argument justifies, in clinical trials of a new drug, reporting statistical summaries of the outcome of therapy rather than, case by case, the best guess of the attending physician about why the result of therapy was what it was for this case. And the same type of argument again justifies using statistical methods for observation on countries or regions even when we could adduce a great deal of special information about each observation which might possibly help explain its "random" component.

Economists must inevitably try to sort out systematic patterns from

random variations in the past—if only because, unassisted, policymakers would do the same thing more naively. In doing so economists will need probabilistic models and statistical methods of inference. Like historians, though, they must accept that a single agreed view of the causal structure of the record they examine will never emerge. Perhaps Hicks's stimulating but unsatisfying book will lead someone to work out a philosophy of inference which confronts both these aspects of economics as science.

References

Burks, Arthur W. *Chance, Cause, Reason: An Inquiry into the Nature of Scientific Evidence*. Chicago and London: Univ. Chicago Press, 1977.

Hicks, John R. *Causality in Economics*. New York: Basic, 1979.

Comments

To Save or Savor: The Rate of Return to Storing Wine

Elizabeth Jaeger

University of Virginia and Freemark Abbey Winery

For $28,000 one can buy a Jaguar XJ6, 308 shares of McDonnell Douglas stock, a one-tenth interest in an Oklahoma gas well, 4½ years of undergraduate life at Stanford, or one bottle of Chateau Lafite, vintage 1806. "That's $1,166.67 an ounce. At this moment gold is worth only $275 an ounce, and you can't smell or drink it" (*Saturday Review of Literature* 1979, p. 38). Nine years ago, $70 could persuade a Paris restaurant to part with this same bottle of liquid gold (Goldwyn 1979, p. 4).

But for balance admire the growth of a slightly more pedestrian vintage, a 1959 Beaulieu estate-bottled Cabernet. Today a case of it may be purchased for $1,200. For the more modest budget, $950 will buy a case of 1970 Beaulieu George Latour Private Reserve Cabernet. All this goes to show that wine, the right wine, in the last 10 years has provided the prudently daring investor with a spectacular return.

The virtues of wine as a liquid asset have been tested recently by Krasker (1979). Krasker cannot reject a hypothesis that wine offers no extraordinary rate of return; the wine investor can expect no greater return than that which accrues to riskless assets. While Krasker finds the concept that an investment in wine outperforms Treasury bills in nominal return "incompatible with the way economists believe most assets are priced," his result seems counter to intuition. Wine as a subcategory of riskless assets seems contrary to wine's very nature.

Wine is a living substance. Its maturation process extends well beyond the bottling day, lasting as long as 20–40 years in the case of some reds, while a few others refuse to retire at 65. But *the* proper age

[*Journal of Political Economy*, 1981, vol. 89, no. 3]
© 1981 by The University of Chicago 0022-3808/81/8903-0010$01.50

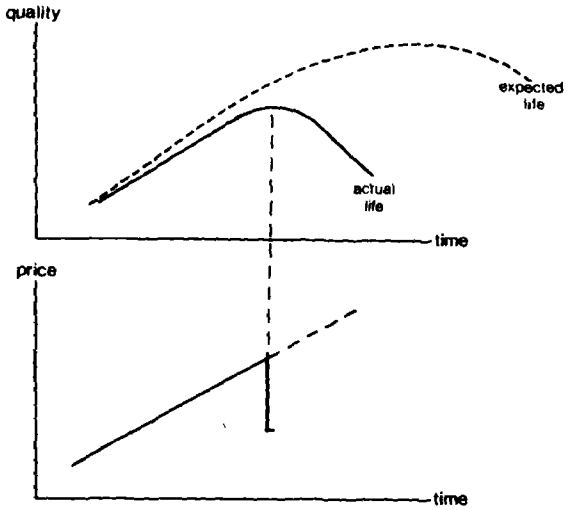


FIG. 1

at which to consume a wine is not some known date but rather a peak which must be discovered by the bulk purchaser through experimentation. When frequently posed with the question when to drink, Harry Waugh, the famous wine writer, would respond, "Just imagine looking around a room full of people and somebody saying to you, 'Tell me at what date each of these individuals is going to die' " (Waugh 1968, p. 1).

This element of uncertainty is reflected in the development of the 1966 vintage French burgundies. These wines were considered years away from full ripeness when, in 1975, out went the cry, "Drink your 1966's!"; they had unexpectedly gone over the hill (Robards 1976, p. 38).

Such events explain some of the sudden downturns in wine prices. Figure 1 traces both the expected and actual maturation of a hypothetical vintage and its price through time. If, at some point in the life of the wine, investors discover that it has peaked prematurely, then they readjust their offer price to reflect this new information. On the other hand, a publicized tasting which revealed much life in a wine of old vintage could be expected to raise its value to the investor.

Quality uncertainty need not, by itself, imply that wine will command a risk premium unless the variance in the return to wine is large relative to the total variance of a portfolio. If wine speculators do not hold a "market" portfolio, such a possibility may exist.

In addition to unanticipated quality changes, wine is affected by cyclical demand fluctuations which can severely affect wine prices. For

example, the 1974–75 recession coincided with a decline in win prices. This covariance with other asset returns is an additional reaso that wine will command a risk premium.

In view of the uncertainty associated with wine's growth and devel opment and of wine's covariance with other assets in a potential portfolio, Krasker's empirical finding that investors receive no risk premium for holding wine is surprising. This paper suggests an alternate method for estimating the wine investor's return.

My approach differs from Krasker's in three important respects. Krasker's data on wine prices are taken from the annual Heublein Wine Auction and cover the 4 years 1973–74, 1974–75, 1975–76, and 1976–77. By limiting his data to these 4 years in particular, he has biased his procedure toward finding a subnormal rate of return. Two years of his sample of 4 were extremely poor from the perspective of the wine industry; in early 1974 and through 1975, newspaper headlines proclaimed that "rivers of surplus wine" existed, reflecting the high inventories of wineries and their distributors. In these 2 years, wine prices, the French especially, fell precipitously below those prices of many preceding years. The negative trend of these figures could be expected to impact heavily on wine returns calculated during this period. My data include the prices determined by the Heublein Auction over 8 years, beginning in 1969 and ending with the 1977 auction. Observations covering 4 more years may offset the dampening effect of the 1974–75 and 1975–76 auctions on the total return to wine storage.

Second, I eliminate an unknown in Krasker's regression: the cost of storing wine. Where his regression estimate of the cost to the wine investor is $16.80 per case per year, I use the cost accounting figure of Freemark Abbey Winery, an annual per case storage cost of $.449. This number is calculated on the basis of the opportunity cost of the storage facility divided by the number of cases. The Freemark Abbey storage cost figure should reflect the cost for the auction participants who, in general, store wine in commercial cellars.

For comparison, however, I estimate the annual cellar costs and the rate of return to wine storage; in this way I have a basis for comparing Krasker's and my results.

Finally, my estimation procedure will differ from Krasker's generalized least-squares process. Limitations on available computer software forced me to employ an alternative estimation procedure which I explain below.

The Model

Following Krasker, I restrict my attention to include only red Bordeaux and California Cabernet Sauvignon produced since 1950.

Thus, I concentrate on those wines which are commonly found to benefit by extended aging, and I exclude those wines which Krasker believed were traded as antiques.[1] An observation is recorded every time a particular vintage wine is traded in consecutive years; then over the year t, $t + 1$, the rate of return to wine i may be represented by $(p_{it+1} - p_{it})/p_{it}$, where p_{it} is the price of wine i 1 year prior to the price observation on the same wine, p_{t+1i}.

Krasker tests the hypothesis that the expected rate of return to storing wine is equal to the rate of return to riskless assets; that is,

$$\frac{Ep_{t+1i} - p_{ti} - \delta}{p_{ti}} - r = \theta = 0, \tag{1}$$

where θ is the risk premium, δ is the cost of wine storage, and r denotes the average rate of return to 3-month Treasury bills for the four quarters which precede the auction. Equation (1) implies the following regression equation:

$$\frac{p_{t+1i} - p_{ti}}{p_{ti}} - r_t = \theta + \delta \frac{1}{p_{ti}} + \frac{u_{t+1i}}{p_{ti}}, \tag{2}$$

where the error term u_{t+1i} is serially uncorrelated as expectations are assumed to be rational, and u_{t+1i}/p_{ti} is assumed to have mean zero and constant variance.

Because the disturbances are correlated within each year, Krasker employed a GLS procedure to derive the estimated equation (standard errors in parentheses):

$$\frac{p_{t+1i} - p_{ti}}{p_{ti}} - r_t = \underset{(.037)}{-.059} + \underset{(.72)}{1.4} \frac{1}{p_{ti}} + \frac{u_{t+1i}}{p_{ti}}, \tag{3}$$

$$R^2 = .03; \text{ SE} = .24.$$

The estimate implies that wine costs \$1.40 per bottle (\$16.80 per case) to store. Although not significantly different from zero, the return differential between wine and Treasury bills is negative.

The Fixed-Effects Model

Because the sample data involve different numbers of observations for each different year, the GLS approach is not straightforward. Krasker does not mention how he estimates the off-diagonal component of the variance-covariance matrix for his GLS estimator, and

[1] Although prestige and publicity over bottle content assuredly guided the purchaser of the 1806 Chateau Lafite, many truly fine wines have been excluded by the arbitrary choice of 1950 for a cutoff point. Nonetheless, I strive to maintain the similarities of these two papers where I can, so I have adopted this same breakpoint.

instead of writing my own seemingly unrelated regression program, I shall consider an alternate method to account for disturbances which are correlated within each year: least-squares estimation with dummy variables. The dummy variable technique exploits the special feature of the combined cross-section/time-series data by explicitly accounting for the "constant" portion of each error term for each given cross section.

As in the GLS procedure, this method presumes that each residual u_{t+1i}/p_{ti} for observations in a given year may be decomposed into two independent components: an individual effect, v_{ti}, and a remainder, ϵ_t. Then, $u_{t+1i}/p_{ti} = \epsilon_t + v_{ti}$, where, in terms of the present example, ϵ_t represents a nonrandom addition to the rate of return to all wine sold in the year t, and v_{ti} becomes the specific classical error term associated with the ith wine, that is,

$$Ev_{ti}v'_{t'j} = \begin{cases} \sigma_v^2, \text{ for } t = t' \text{ and } i = j \\ 0, \text{ for } t \neq t' \text{ or } i \neq j. \end{cases}$$

The ϵ_t component of the rate of return can be incorporated into Krasker's regression equation by the introduction of dummy variables to represent given years. In this way, each year is characterized by its own special intercept. The regression equation is then a fixed-effects model, using Scheffé's (1959) terminology.

For ease of exposition, I will denote the dependent variable as y_{ti}. The vector $\mathbf{y}_k$ ($k = 1, \ldots, m$) will contain all observations on the dependent variables for year k. The variable $1/p_{ti}$ will be denoted z_{ti}, and $\mathbf{z}_k$ will be the vector of observations on z_{ti} for year k. The vector $\mathbf{v}_k$ ($k = 1, \ldots, m$) will contain all of the observations in v_{ti} for year k. Finally, $\mathbf{i}_k$ ($k = 2, \ldots, m$) represents a vector with entries of one in the elements associated with the kth year and zeroes elsewhere. Stacking the vectors, let

$$\mathbf{y} = \begin{bmatrix} \mathbf{y}_1 \\ \vdots \\ \mathbf{y}_m \end{bmatrix}, \quad \mathbf{z} = \begin{bmatrix} \mathbf{z}_1 \\ \vdots \\ \mathbf{z}_m \end{bmatrix}, \quad \text{and } \mathbf{v} = \begin{bmatrix} \mathbf{v}_1 \\ \vdots \\ \mathbf{v}_m \end{bmatrix}.$$

The relationship to be estimated from the combination of time-series and cross-section data can then be expressed in the form

$$\mathbf{y} = \mathbf{c} + \beta_2\mathbf{i}_2 + \ldots + \beta_m\mathbf{i}_m + \delta_\mathbf{z} + \mathbf{v}, \tag{4}$$

where (c, β_2, β_3, . . . , β_m, δ) is a vector of unknown parameters. Equation (4) can be estimated efficiently with ordinary least squares. The parameter estimates can be used to estimate the quantity (c + $[1/m] \sum_{i=2}^{m} \beta_i$); the risk premium that one would expect to accrue to a wine is randomly selected from the total sample.

The application of this procedure to the wine data indeed reveals a premium to the wine investor with one qualification. Running four

TABLE 1

THE RETURN TO WINE STORAGE

	Regressions Using 4-Year Data (N = 140)		Regressions Using 8-Year Data (N = 199)	
	Regression Estimate of Storage Cost	Storage Cost Estimate $.449	Regression Estimate of Storage Cost	Storage Cost Estimate $.449
Average premium to wine investor	−.01367	.03812	.08542	.16614
Standard error (average premium)	.03399	.0661	.03379	.03296
t-statistic	−.4021	5767	2.5279	5.0392
δ	17.25	.	16.60	...
(SE)	(8.76)		(5.13)	
R^2	.29	.27	.41	.38
Standard error	2415	.2439	2532	2587
Mean dependent variable	0182	0168	.1036	.1068

separate regressions, I have distinguished between the effects on the rate of return of (1) increasing the period of observations from 4 to 8 years and (2) estimating (on the right-hand side) versus incorporating (on the left-hand side) the storage cost in the regression equation. Specifically, I run two regressions on equation (4), one using Krasker's 140 observations taken over 4 years and the second using 199 observations from an 8-year time span. In an identical fashion, I run two more regressions on the following version of equation (4),

$$\mathbf{y} - \delta\mathbf{z} = c + \beta_2\mathbf{i}_2 + \ldots + \beta_m\mathbf{i}_m + \mathbf{v}, \tag{5}$$

where δ = .449, the Freemark Abbey storage cost estimate. Using the estimates for $(\hat{c}, \hat{\beta}_2, \ldots, \hat{\beta}_m)$, I have calculated $\hat{c} + (1/m)\sum_{i=2}^{m}\hat{\beta}_i$ together with its standard error. The results appear in table 1.

Not surprisingly, one estimate confirms Krasker's findings that the expected return to wine lies beneath the return to Treasury bills. This regression utilized Krasker's data over the period 1973–77, where 1973–75 were very poor performance years for the wine industry. And like Krasker's result, the return is biased downward by the unrealistic storage cost estimate produced by the regression: an annual $17.25 per case. The influence of this estimate can be seen in the second regression, where the cellar cost is accounted for using the Freemark Abbey winery estimate; a 3.8 percent premium to oenophiles separates the wine and Treasury investment. However, as in Krasker's study, neither premium is significantly different from zero.

Expanding the data to cover the 8-year span (1969–77) greatly

instead of writing my own seemingly unrelated regression program, I shall consider an alternate method to account for disturbances which are correlated within each year: least-squares estimation with dummy variables. The dummy variable technique exploits the special feature of the combined cross-section/time-series data by explicitly accounting for the "constant" portion of each error term for each given cross section.

As in the GLS procedure, this method presumes that each residual u_{t+1i}/p_{ti} for observations in a given year may be decomposed into two independent components: an individual effect, v_{ti}, and a remainder, ϵ_t. Then, $u_{t+1i}/p_{ti} = \epsilon_t + v_{ti}$, where, in terms of the present example, ϵ_t represents a nonrandom addition to the rate of return to all wine sold in the year t, and v_{ti} becomes the specific classical error term associated with the ith wine, that is,

$$Ev_{ti}v'_{t'j} = \begin{cases} \sigma_v^2, \text{ for } t = t' \text{ and } i = j \\ 0, \text{ for } t \neq t' \text{ or } i \neq j. \end{cases}$$

The ϵ_t component of the rate of return can be incorporated into Krasker's regression equation by the introduction of dummy variables to represent given years. In this way, each year is characterized by its own special intercept. The regression equation is then a fixed-effects model, using Scheffé's (1959) terminology.

For ease of exposition, I will denote the dependent variable as y_{ti}. The vector $\mathbf{y}_k$ ($k = 1, \ldots, m$) will contain all observations on the dependent variables for year k. The variable $1/p_{ti}$ will be denoted z_{ti}, and $\mathbf{z}_k$ will be the vector of observations on z_{ti} for year k. The vector $\mathbf{v}_k$ ($k = 1, \ldots, m$) will contain all of the observations in v_{ti} for year k. Finally, $\mathbf{i}_k$ ($k = 2, \ldots, m$) represents a vector with entries of one in the elements associated with the kth year and zeroes elsewhere. Stacking the vectors, let

$$\mathbf{y} = \begin{bmatrix} \mathbf{y}_1 \\ \vdots \\ \mathbf{y}_m \end{bmatrix}, \quad \mathbf{z} = \begin{bmatrix} \mathbf{z}_1 \\ \vdots \\ \mathbf{z}_m \end{bmatrix}, \text{ and } \mathbf{v} = \begin{bmatrix} \mathbf{v}_1 \\ \vdots \\ \mathbf{v}_m \end{bmatrix}.$$

The relationship to be estimated from the combination of time-series and cross-section data can then be expressed in the form

$$\mathbf{y} = \mathbf{c} + \beta_2\mathbf{i}_2 + \ldots + \beta_m\mathbf{i}_m + \delta_\mathbf{z} + \mathbf{v}, \tag{4}$$

where (c, β_2, β_3, . . . , β_m, δ) is a vector of unknown parameters. Equation (4) can be estimated efficiently with ordinary least squares. The parameter estimates can be used to estimate the quantity (c + $[1/m] \sum_{i=2}^{m} \beta_i$); the risk premium that one would expect to accrue to a wine is randomly selected from the total sample.

The application of this procedure to the wine data indeed reveals a premium to the wine investor with one qualification. Running four

TABLE 1

THE RETURN TO WINE STORAGE

	Regressions Using 4-Year Data (N = 140)		Regressions Using 8-Year Data (N = 199)	
	Regression Estimate of Storage Cost	Storage Cost Estimate $.449	Regression Estimate of Storage Cost	Storage Cost Estimate $.449
Average premium to wine investor	−.01367	.03812	.08542	.16614
Standard error (average premium)	.03399	.0661	.03379	.03296
t-statistic	−.4021	.5767	2.5279	5.0392
δ	17.25		16.60	. .
(SE)	(8.76)		(5.13)	
R^2	.29	.27	.41	.38
Standard error	.2415	.2439	.2532	.2587
Mean dependent variable	.0182	.0168	.1036	.1068

separate regressions, I have distinguished between the effects on the rate of return of (1) increasing the period of observations from 4 to 8 years and (2) estimating (on the right-hand side) versus incorporating (on the left-hand side) the storage cost in the regression equation. Specifically, I run two regressions on equation (4), one using Krasker's 140 observations taken over 4 years and the second using 199 observations from an 8-year time span. In an identical fashion, I run two more regressions on the following version of equation (4),

$$\mathbf{y} - \delta\mathbf{z} = \epsilon + \beta_2\mathbf{i}_2 + \ldots + \beta_m\mathbf{i}_m + \mathbf{v}, \tag{5}$$

where δ = .449, the Freemark Abbey storage cost estimate. Using the estimates for $(\hat{\epsilon}, \hat{\beta}_2, \ldots, \hat{\beta}_m)$, I have calculated $\hat{\epsilon} + (1/m)\sum_{i=2}^{m}\hat{\beta}_i$ together with its standard error. The results appear in table 1.

Not surprisingly, one estimate confirms Krasker's findings that the expected return to wine lies beneath the return to Treasury bills. This regression utilized Krasker's data over the period 1973–77, where 1973–75 were very poor performance years for the wine industry. And like Krasker's result, the return is biased downward by the unrealistic storage cost estimate produced by the regression: an annual $17.25 per case. The influence of this estimate can be seen in the second regression, where the cellar cost is accounted for using the Freemark Abbey winery estimate; a 3.8 percent premium to oenophiles separates the wine and Treasury investment. However, as in Krasker's study, neither premium is significantly different from zero.

Expanding the data to cover the 8-year span (1969–77) greatly

improves the calculated return to wine. A low dividend of 8.5 percent reflects the high estimated cellar cost of $16.60. But wine outperforms Treasury bills by an impressive 16.6 percent for the commercial establishment for whom the Freemark Abbey cost gauge applies. Both premiums are significantly different from zero.

Reinterpreting Krasker's Regression Equation

A somewhat disturbing question which I have yet to address is why Krasker's and my regression equation with unrestricted storage costs, that is

$$\frac{p_{t+1i} - p_{ti}}{p_{ti}} - r_t = \theta + \delta \frac{1}{p_{ti}} + \frac{u_{t+1i}}{p_{ti}},$$

give the implausible estimate, attributed to $1/p_{ti}$, of $16.60 per case annually. However, note that $16.60/p_{ti}$ is a quantity on the order of magnitude of a rate of return, since the average value of p_{ti} over the 8-year period is $410. An alternative interpretation of this result is that $\delta(1/p_{ti})$ does not represent annual storage costs but rather an as yet unexplained portion of the rate of return to wine storage.

If the premium of the wine investor can be segregated into two effects, one which accrues to the wine regardless of price and a second effect related to its nominal price, then $\theta + (\delta/p_{ti})$ will represent the total return to storing wine. In this case δ/p_{ti} captures the premium as a function of price. Krasker's positive estimate for δ, interpreted in this light, implies a higher premium to lower-priced wines. The alternative postulate that high-priced wines receive the better rate of return would be supported by a negative δ.

The possibly surprising conclusion that lower-priced wines receive higher premiums than their counterparts can be explained by risk. Wines attached to smaller nominal price tags for some reason entail greater risk and return for the investor. Thus price stands as a proxy for risk, where risk can be measured by the variance of the rate of return. The validity of this explanation for Krasker's "cost" estimate relies on the variance of the rate of return of high- and low-priced wines, and support for this interpretation of the regression equation would be gained by a measurably higher rate of return variance in the low-price group. Ordering the 199 observations from the 8 auction years by nominal price, I have computed the mean rate of return and variance for the 65 highest- and 66 lowest-priced wines. While wines in the high price range offered the investor an average rate of return of .03 (3 percent) with a variance of .051, the lowest one-third provided a mean rate of return of .3587 (36 percent) with a variance

more than twice as large as the highest third, .1162. Performing an F-test, I can reject the hypothesis that the two variances are equal at the 95 percent level of significance, since the value of the F-statistic is 2.28.

Repeating the procedure using 140 observations from the 4 auction years considered by Krasker yields similar results. The 47 highest-priced wines provide a mean return of −.0267, with a variance of .044; the 47 lowest-range wines earned an average return of .1843 (18 percent) with a variance of .086, double that of the high-priced wines. Again, based on an F-statistic of 1.95, I can reject the hypothesis that the two variances are equal at the 95 percent level of significance.

Apparently, then, wine has an even higher return than allowed earlier in this paper, since previous calculations did not incorporate the return associated with a wine's own price but rather attributed that portion of the rate of return to cost. Using the cellar cost statistic computed by Freemark Abbey Winery, we can infer from Krasker's regression that

$$\frac{p_{t+u} - p_{tu}}{p_{tu}} - r_{tu} - .449\frac{1}{p_{tu}} = \left(\theta + \delta\frac{1}{p_{tu}}\right) + \frac{u_{t+u}}{p_{tu}}. \qquad (6)$$

The amended risk premium to storing wine for the 8-year period, derived by computing $\hat{c} + (1/m)\sum_{i=2}^{m}\beta + (\hat{\delta}/\bar{p})$, with $\bar{p} = \$410$, is .1239 with a standard error of .0672.

Before proposing an intuitive explanation for the greater risk associated with lower-priced wines, it should be emphasized that "lower priced" is relative to this sample of expensive premium wines. By no means do I mean to imply that risk is associated with less costly wines in general; consistency (i.e., zero risk) is the by-line of Ernest and Julio Gallo's jug wines.

One plausible interpretation of the result is that the most expensive wines, heavily represented by the older vintages, offer greater certainty to the investor because of their age. Their more youthful counterpart, with a shorter track record, provides the wine buyer with unrealized potential, and thus risk, at a discounted price.

Conclusion

In reestimating the rate of return premium to wine storage using a larger data set, I have reversed Krasker's findings that wine speculation yields no greater return than do Treasury bills. In addition, I have shown that the use of an actual storage cost measurement allows reinterpretation of Krasker's estimated "storage cost" as part of the return to wine. Adding this return to Krasker's original estimates also produces a positive premium to wine storage.

References

Goldwyn, Craig. *Chicago Tribune* (June 18, 1979).

Krasker, William S. "The Rate of Return to Storing Wines." *J.P.E.* 87, no. 6 (December 1979): 1363–67.

Robards, Terry. *The New York Times Book of Wine*. New York: New York Times Book Co., 1976.

Saturday Review of Literature. "Liquid Assets, Wine as an Investment" (July 7, 1979).

Scheffé, Henry. *The Analysis of Variance*. New York: Wiley, 1959.

Waugh, Harry. *The Changing Face of Wine: An Assessment of Some Current Vintages*. London: Wine and Spirit Publications, 1968.

On the Relationship between Commodity Price Changes and Factor Owners' Real Positions

James Cassing

Australian National University

Recent contributions to trade theory have utilized the generalized variable proportions model in order to identify "natural friends" and "natural enemies" of particular relative commodity-price changes (see, e.g., Jones and Scheinkman [1977] and Jones [1979, chap. 8] for a relevant survey). In the case where the number of factors equals the number of commodities, it is now known that every factor has a natural enemy in some commodity in the sense that any factor's reward is inversely related to some commodity's price. Also, it is known that, in general, a given factor need not have a natural friend in the sense that the factor's reward may not increase more than proportionately to any commodity's price increase.

However, this result should not be interpreted to imply that in the $n \times n$ case some individual whose endowment consists only of a given factor may not be able to increase his real income through an appropriate commodity-price rise. On the contrary, when we take into account an individual's expenditure pattern, it can be shown that any individual can improve his real position through some commodity-price rise.

In order to see this, consider the standard $n \times n$ variable proportions trade model. Let Θ denote the $n \times n$ distributive share matrix, w the $n \times 1$ vector of factor rewards, and p the $n \times 1$ vector of commodity prices. Using "$\hat{}$" to denote a proportionate change, we have the usual price equation

$$\Theta'\hat{w} = \hat{p}. \tag{1}$$

This paper has benefited from the helpful comments of Peter Warr and an anonymous referee.

[*Journal of Political Economy*, 1981, vol. 89, no. 3]
© 1981 by The University of Chicago. 0022-3808/81/8903-0011$01.50

We assume that Θ' is nonsingular so that

$$\hat{w} = \Theta'^{-1}\hat{p}. \tag{2}$$

Jones and Scheinkman show with an example that for any factor i the ith row of Θ'^{-1} may contain elements all less than unity.

But the change in an individual's real position also depends upon expenditure patterns. Suppose that agent l's endowment consists only of the ith factor. Denote by ϕ^l_{ik} the lth agent's expenditure share of income on commodity k. Then the real income effect for individual l of a change in any commodity k's price is given by

$$\hat{w}_i - \phi^l_{ik}\hat{p}_k. \tag{3}$$

Substitution from (2) yields the real income effect

$$(\Theta'^{-1}_{ik} - \phi^l_{ik})\hat{p}_k, \tag{4}$$

where Θ'^{-1}_{ik} is the $i - k$th element of Θ'^{-1}.

The real position of agent l improves if the term in parentheses in equation (4) is positive. We know that this term is negative for some k, since all factors have natural enemies. But the sum of these coefficients over all k is zero, since Θ' is row stochastic, so that Θ'^{-1} has unit row sums, and since all income is spent. That is,

$$\sum_{k=1}^{n} \Theta'^{-1}_{ik} - \sum_{k=1}^{n} \phi^l_{ik} = 1 - 1 = 0. \tag{5}$$

Therefore, since the summation has at least one negative term, it must have a positive term as well. The real position of any individual can be improved through an appropriate commodity-price rise.

As it stands, this result is not the exact counterpart of the "existence of a natural enemy" result which depends on only the technology. A natural friend would be a commodity for which a price rise induces an improvement in a factor's real wage, independent of the expenditure shares of its different owners used to compute each owner's price deflator. Rather, we show here only that in the $n \times n$ case for each individual l whose endowment consists only of a given factor, there exists a commodity such that, when this commodity price rises, the real wage of that factor, when deflated by l's price index, rises. Thus every individual is a proponent of some—not necessarily the same—commodity-price increase. And, of course, if all owners of a single factor happen to display the same expenditure shares, then there would be a sense in which every factor can improve its real position through some commodity-price rise.

References

Jones, Ronald W. *International Trade: Essays in Theory*. Amsterdam: North-Holland, 1979.

Jones, Ronald W., and Scheinkman, José A. "The Relevance of the Two-Sector Production Model in Trade Theory." *J.P.E.* 85, no. 5 (October 1977): 909–35.

Confirmations and Contradictions

Stochastic Implications of the Life Cycle–Permanent Income Hypothesis: Evidence for the U.K. Economy

Vince Daly and George Hadjimatheou

Kingston Polytechnic

In a previous issue of this *Journal*, Hall (1978) presented a simple life cycle–permanent income model of the consumption decision according to which consumption expenditures are predicted to follow a random walk with trend; that is,

$$C_t = \lambda C_{t-1} + \epsilon_t, \qquad \lambda > 1, \tag{1}$$

where ϵ_t is white noise. Hall's stated purpose in deriving this equation was to provide a testable implication of the pure life cycle–permanent income hypothesis which avoided the econometric problems associated with a consumption function in which measured income is included as an explanatory variable. The testable implication is that, apart from this period's consumption expenditure, no variable observable in this or in earlier periods should show any predictive power for next period's consumption expenditure. For U.S. quarterly data 1948(1)–1977(1), Hall found this to be the case, with the exception of movements in the real value of corporate stock.[1] Given the importance of the hypothesis it would seem worthwhile to examine the extent to which it is substantiated by alternative economic histories, and so we present here the results obtained from U.K. quarterly data in the postwar period.

[1] In Hall's paper the significance of the corporate stock variable is not considered as contradictory to the tested hypothesis when the model is modified to allow for a lagged adjustment of consumption to a change in permanent income. The implicit assumption is that, given the random walk nature of corporate stock values, changes in them are expected to be correlated with changes in permanent income.

[*Journal of Political Economy*, 1981, vol. 89, no. 3]
© 1981 by The University of Chicago. 0022-3808/81/8903-0012$01.50

The data are taken from *Economic Trends—Annual Supplement 1980*, and, following Hall's example, variables were, where appropriate, measured per capita and deflated by the implicit price deflator of nondurable consumption expenditure, the latter being the measure of consumption adopted in the original study. Since the test procedure involved searching for variables which could supplement the predictive power of equation (1), a large number of regression results were generated which we do not report below; we simply note that unemployment, the inflation rate, and the rate of interest, all of which have been proposed on occasion as determinants of C_t, did not contribute significantly to the explanatory power of a regression equation containing C_{t-1}.

Our substantive result is that regression specifications which include lagged values of disposable income or liquid assets, or the period-to-period changes in these variables, and regressions which include consumption expenditures lagged by more than one period can significantly outperform the simpler specification in which consumption expenditure is predicted only by its own lagged value.[2] Thus some doubt is cast on both the validity of the pure life cycle–permanent income hypothesis, in the case of the United Kingdom, and the policy prescriptions that might follow from an acceptance of the hypothesis.

It can be seen in the Appendix that both consumption expenditure lagged more than once and lagged income appear with coefficients which are significantly different from zero. Furthermore, for the shorter period 1964–78 for which data on liquid assets are available, changes in the latter in the recent past seem to contribute significantly to the explanatory power of the regression. In all cases the size of the F_1-statistic suggests that the hypothesis that the additional variables do not contribute to the regression should be rejected.[3]

Finally, repeated estimation for different subperiods of the simple regression of C_t on C_{t-1} indicates that the coefficient on the latter

[2] The relevance of liquid assets in explaining consumers' expenditure has been suggested in a number of recent studies; see, e.g., Townend 1976

[3] When 10 observations of each sample are retained for a postsample parameter stability test of the alternative specifications, the χ^2 (Davidson et al. 1978, p. 674) and Chow's F-test suggest that in all cases the hypotheses of postsample parameter stability and of no structural change should not be rejected. What is more important, however, for the present study is that the specification which allows for the influence of changes in liquid assets seems to provide more accurate forecasts than the pure life cycle–permanent income hypothesis. The superiority of the alternative specification is also maintained when stability is tested for post-1972 data—a critical period for econometric forecasting, coinciding with a dramatic increase in the U.K. saving ratio (Hadjimatheou 1979). For this period, however, the value of χ^2 suggests that the hypothesis of postsample parameter stability should be rejected for both specifications while the F-value is still not statistically significant.

variable is unstable. More specifically, as can be seen from the following values of $\hat{\lambda}$, the greater the time span of the sample, moving backward toward the less recent past, the higher the level of the coefficient: (i) 1973(1)–1978(4), 1.00125; (ii) 1964(1)–1978(4), 1.00374; (iii) 1956(1)–1978(4), 1.00413; (iv) 1964(1)–1972(4), 1.00483; (v) 1956(1)–1972(4), 1.00497 (Chow test: iii and v, $F(24,66) = 3.14$; ii and iv, $F(24,34) = 2.21$). In the context of a more conventional consumption function this is not surprising, given the marked downward trend of the average propensity to consume in the United Kingdom, especially in the 1970s. This observation, however, does not seem to lend support to Hall's statement that constancy of λ is expected to "be a good approximation, at least over a decade or two" (Hall 1978, p. 975).

In conclusion, the present evidence seems to suggest rejection of the pure life cycle–permanent income hypothesis in the U.K. case. Furthermore, the latter cannot explain the revealed differences between the two countries. It may be tempting to suggest that English people are less "rational" than Americans, but such a pronouncement would certainly need an explanation.

What is probably needed is a more general theory of the consumption decision which encompasses the different experience of the two countries as special cases. It may be that binding financial constraints, changes in the age structure and the distribution of income, as well as changes in interest rates and the composition of wealth, are too important to be ignored. The importance of some of these factors has, after all, been recognized, either explicitly or implicitly, by the life cycle and/or permanent income hypotheses.

Appendix

Dependent Variable is C

1956(1)–1978(4), $N = 91$:

$$\underset{(1055)}{1.00413C_{t-1}}, \tag{A1}$$

$$\bar{R}^2 = 0.994, \quad S = 2.103, \quad SSR = 398.2.$$

$$\underset{(7.9)}{0.819C_{t-1}} + \underset{(2.1)}{0.2197C_{t-2}} + \underset{(2.9)}{0.1201Y_{t-1}} - \underset{(3.6)}{0.1492Y_{t-2}}, \tag{A1.1}$$

$$\bar{R}^2 = 0.995, \quad S = 1.986, \quad SSR = 343.1, \quad F_1(3,87) = 4.66.$$

1964(1)–1978(4), $N = 59$:

$$\underset{(780)}{1.00374C_{t-1}} \tag{A2}$$

$$\bar{R}^2 = 0.971, \quad S = 2.441, \quad SSR = 346.0.$$

$$\underset{(4.1)}{0.4971C_{t-1}} + \underset{(1.67)}{0.2353C_{t-2}} + \underset{(2.3)}{0.273C_{t-3}} + \underset{(1.92)}{0.0895\Delta Y_{t-1}} + \underset{(5.2)}{0.115\Delta LA_{t-1}}, \tag{A2.1}$$

$$\bar{R}^2 = 0.9884, \quad S = 1.881, \quad SSR = 191.0, \quad F_1(4,54) = 10.96.$$

$$\underset{(4.8)}{0.604C_{t-1}} + \underset{(3.6)}{0.442C_{t-2}} + \underset{(1.60)}{0.0831Y_{t-1}} - \underset{(1.79)}{0.105Y_{t-2}} + \underset{(3.8)}{0.0886LA_{t-1}} - \underset{(4.0)}{0.0926LA_{t-2}}, \tag{A2.2}$$

$$\bar{R}^2 = 0.988, \quad S = 1.96, \quad SSR = 204, \quad F_1(5,53) = 7.38.$$

Numbers in parentheses are t-statistics; S = standard error of estimate, SSR = sum of the squares of the residuals, and F_1 is a test for the significance of additional parameters; $\Delta Y_{t-1} = Y_{t-1} - Y_{t-2}$ and $\Delta LA_{t-1} = LA_{t-1} - LA_{t-2}$; LA = selected liquid assets of the personal sector as given in *Financial Statistics*.

References

Davidson, James E. H.; Hendry, David F.; Srba, Frank; and Yeo, Stephen "Econometric Modelling of the Aggregate Time-Series Relationship between Consumers' Expenditure and Income in the United Kingdom." *Econ. J.* 88 (December 1978): 661–92.

Hadjimatheou, George. "An Explanation of the Decline in the Average Propensity to Consume in the U.K." *British Rev. Econ. Issues* 2 (November 1979): 1–13.

Hall, Robert E. "Stochastic Implications of the Life Cycle–Permanent Income Hypothesis: Theory and Evidence." *J.P.E.* 86, no. 6 (December 1978): 971–87.

Townend, J. C. "The Personal Saving Ratio." *Bank of England Q. Bull.* 16 (March 1976): 53–73.

Miscellany

Why There Are No Risk Preferrers

Somewhat more than 100 years ago, Smith's celebrated diamond-water paradox was solved by the observation that however valuable water might be to someone who had none, if it was produced and sold in unlimited quantities at a price p consumers would adjust their rate of consumption so that its marginal value to them was at most p. The same argument implies that behavior under uncertainty depends not merely on the actor's utility function but also on the cost of producing (and absorbing) risk. However fond I may be of uncertainty, I will not buy it at a high price (by choosing a profession whose returns have high variance but low expected value) when I can get it at a low one (in Las Vegas).

Risk can be produced inexpensively; existing sellers in Las Vegas and elsewhere charge only a few percent of the expected value of the gamble, part of which must be regarded as payment not for the risk itself but for the palatial surroundings in which it is produced. Hence, absent special circumstances (such as a government monopoly of risk production), we may expect individuals in equilibrium to exhibit at most a small degree of risk preference, for the same reason that they exhibit at most a small marginal utility for water.

Insurance, unlike risk, is inherently costly to produce because of the problems of moral hazard. Moral hazard is a major cost of insurance but a minor cost of gambling because a gambling game, being created for the purpose of producing risk, can be designed to almost eliminate the ability of participants to spend resources on influencing the outcome. The seller of insurance faces a preexisting situation in which it is likely that the buyer (of, say, unemployment insurance) can and will affect the odds by his actions. Hence insurance is costly to produce and individuals in equilibrium may exhibit a substantial degree of risk avoidance.

DAVID FRIEDMAN

Virginia Polytechnic Institute

Journal of Political Economy, 1981, vol. 89, no. 3]
© 1981 by The University of Chicago 0022-3808/81/8903-0013$01.50

Book Reviews

The Regulatory Process and Labor Earnings. By RONALD G. EHRENBERG. New York: Academic Press, 1979. Pp. xx+204. $15.00.

The irony is that this study was not published in the *Bell Journal of Economics.* Ehrenberg's argument is that the wage rates of workers in the New York Telephone Company are above those of other equally qualified workers and that therefore the New York Public Service Commission should not allow further wage increases to be passed on as higher telephone rates. As I read through his argument I wondered how Ehrenberg would have treated a request to cover increased publication costs by the editors of the *Bell Journal.* A related tactic did not escape notice by the leaders of the Communication Workers of America (CWA), however, who were on the other side of the regulatory proceeding in which Ehrenberg participated. In addition to implying that Ehrenberg was a mere hired gun for Alfred Kahn (then chairman of the New York State Public Service Commission), according to Ehrenberg, ". . . political pressure was put on my employer, the New York State School of Industrial and Labor Relations (ILR) at Cornell University, by the New York State AFL-CIO, which cancelled a scheduled conference at Cornell, implicitly suggesting that it would be wise if ILR faculty did not get involved in similar cases in the future" (pp. 147–48). The hero of this part of the story was apparently Bob McKersie (then dean of the ILR school but now at MIT) who refused to take institutional responsibility for an individual faculty member's research conclusions.

I think most economists will find this a fascinating study. In fact, I would urge any economist looking for a good read while on holiday to take this (relatively inexpensive) book along. I took it and a pretty good paperback for a week on Martha's Vineyard (spending a whole summer there is apparently a prerequisite for becoming president of the AEA) and actually finished Ehrenberg's book first. For one thing, it is well written. Regression equations and the like are kept in an appendix. As a consequence the book would make excellent supplementary reading for graduate or undergraduate courses in labor economics, industrial organization, public policy, or applied econometrics.

For another thing, the book tells a coherent story about a complete policy-oriented research project from start to finish. The 1976 inception of Ehrenberg's study resulted from an unprecedented request by the New York Telephone Company for a rate increase to cover future, yet unnegotiated,

[*Journal of Political Economy*, 1981, vol. 89, no. 3]
© 1981 by The University of Chicago

increases in wages and benefits. The conclusion of the study comes a year later when the Public Service Commission declines to accept Ehrenberg's logic and therefore finds his empirical calculations inappropriate to the decision at hand. In the meantime, however, the union-management negotiations had been completed under the shadow of a possibly adverse decision on cost pass-through, and, I would bet, this was just the purpose for which Ehrenberg's study was initiated by the Public Service Commission staff in the first place.

Ehrenberg's study is also coherent in another dimension. It reflects a remarkable breadth of substantive argument over a range of subjects covering labor law, the determinants of unionization and bargaining structure, as well as some very convincing empirical economics. For labor economists there are some special treats. Ehrenberg tells us that he billed the Public Service Commission staff for 40 days of consulting time (at an undisclosed rate), while National Economic Research Associates billed New York Telephone for 87 days of consulting time (at an undisclosed, but higher, rate). I am not sure what this is supposed to imply for public policy, but it does raise a larger issue. As the real academic incomes of economists have fallen throughout the last decade (I recently calculated that junior faculty taking jobs at my university this year will receive real salaries 22 percent below the starting salaries 12 years ago), their efforts have naturally turned toward the more practical and remunerative. One cannot help wondering what effect this trend will have on the profession's research output as this income squeeze intensifies during the decade of falling college enrollments that we are entering

The core of Ehrenberg's study is his empirical comparison of wage rates and fringe benefits at New York Telephone with the wage rates of other workers. Ehrenberg deliberately chooses "other workers" to be other nonunion workers with similar skills, a decision the Public Service Commission ultimately found inappropriate. Wage comparisons are first made using several surveys from the Bureau of Labor Statistics and other sources. Since worker human capital is difficult to control in these comparisons, standard wage regressions from the 1/100 sample of the 1970 Census data for New York State are also computed. The telephone company is so large an employer, and New York so large a state, that this procedure produces relatively precise wage-differential estimates. Taken together, I found Ehrenberg's empirical case convincing, especially when bolstered with the extraordinarily low relative quit rates at New York Telephone. In sum, I think there is little doubt that wage rates at New York Telephone had by 1976 been increased above market-clearing levels and that many of the employees of this company were earning economic rents. The question that naturally arises, therefore, is what attitude the New York Public Service Commission should have taken toward this stylized fact on the occasion of a rate-increase request.

The administrative law judge in this regulatory proceeding recommended that this "fact" be ignored because to do otherwise would be to interfere in the collective-bargaining process in violation of federal labor law. Thus, the administrative law judge concluded that whether to use Ehrenberg's wage comparisons was mainly a "legal issue." As anyone who has consulted with lawyers will understand, a "legal issue" is primarily one for which there is no accepted answer. In fact, Ehrenberg shows that as a matter of law the administrative law judge was simply incorrect on this point and the Public Service Commission concluded similarly. Despite the commission's conclusion that disallowing excessive labor-cost increases was legally permissible, however, the commission explicitly rejected Ehrenberg's wage comparisons as inappropriate.

In effect, Ehrenberg compared the average wage rate in a unionized monopoly firm against the average of all other wage rates, union and nonunion. Since the majority of all workers are nonunion, this is primarily a comparison of the wage rates of unionized workers in a monopoly firm against the wage rates of nonunion workers. The commission argued that the average wage rate of the workers at New York Telephone should have been compared against the average wage rate of other unionized workers. Since there is an empirically established overall union/nonunion wage differential, this comparison would essentially have uncovered whether the union/nonunion wage differential was systematically greater in a particular regulated-monopoly industry. If this had been the case, the commission apparently would have been prepared to disallow further labor-cost increases as within the scope of its mandate to regulate monopoly power but not union power.

Ehrenberg really offers no convincing argument against the commission's position. Moreover, although he does not offer the necessary empirical work to test the appropriate factual issues, I suspect that such a study would not demonstrate that union/nonunion wage differentials were higher among unionized employees of monopoly firms than among other unionized workers. Historically, for example, union/nonunion wage differentials have been highest among construction workers and mine workers, both relatively competitive industries.

The difficulty here is that the commission's position essentially takes the presence of unionization as exogenous and asks merely whether the presence of regulated monopoly exacerbates union wage pressures. This position ignores the really important question of whether an industry might not have become unionized *because* it is a regulated monopoly. In this case the presence of the union is endogenous to the regulatory process, and the appropriate wage comparison is akin to the one Ehrenberg offers. The anecdotal evidence he presents of the evolution of the Bell System bargaining structure as well as the well-known anecdotal evidence regarding the trucking and railroad industries and the history of the Davis-Bacon Act in contract construction all support this interpretation. As Ehrenberg's experience clearly demonstrates, however, making this case before a regulatory commission is likely to be a thankless task, and perhaps not very remunerative either.

ORLEY ASHENFELTER

Princeton University

Law, Legislation and Liberty. A New Statement of the Liberal Principles of Justice and Political Economy. Vol. 3: *The Political Order of a Free People.* By FRIEDRICH A. HAYEK

Chicago: University of Chicago Press, 1979. Pp. 243. $14.00; London and Henley, Routledge & Kegan Paul, 1979. Pp. 243. £5.95.

> But to avoid this [destroying our civilization] we must shed the illusion that we can deliberately create the future of mankind. This is the final conclusion of the forty years which I have devoted. . . .

I want to thank Kaj Areskoug, Fritz Machlup, Ingo Walter, and Lee Wohlfert for comments on an earlier draft of this review.

These dramatic words end the third and last volume of *Law, Legislation and Liberty: A New Statement of the Liberal Principles of Justice and Political Economy* (hereafter cited as *LLL*) by Nobel laureate Friedrich A. Hayek. The recently published volume, *The Political Order of a Free People*, analyzes the undermining of democratic principles and liberty in parliamentary democracies and concludes with an outline of a new constitutional framework.

The message of Hayek's voluminous work since World War II[1] is, to put it briefly, that the current political system of Western democracies is bound to cause increasing coercion of individuals by governments, private groups, and firms, and that governmental decisions will increasingly be contrary to the opinions of the majority. He argues in volume 3 that the unlimited powers (sovereignty) of governments—combined with the belief and expectation that governments can direct production, consumption, and the distribution of income to maximize social welfare—are the necessary and sufficient conditions for this development. The constitutional proposal is designed to subject the government to substantial constraints in its use of coercive powers.

The previous two volumes of *LLL* (Hayek 1973, 1976*b*) provide the foundation for the analyses of the political process in volume 3, but the last can be read independently. *Rules and Order* analyzes the role of rules and law in a "free" society—one in which no individual can be coerced to act in accordance with the will of others. Law must represent general principles of just conduct on which there is common agreement. Coercion must be used only to enforce such general principles applicable to all individuals. *The Mirage of Social Justice* discusses the limits of reason and rationality in social design and the impossibility of defining social justice. Some arguments in these two volumes will be referred to below to clarify the discussion of volume 3.

The "Bargaining" Democracy and Logrolling

Hayek uses the concept of the bargaining democracy to describe a system of representative assemblies within which different interest groups can trade in vote support for each other's proposals. In the theory of public finance, this is called logrolling[2] and is generally regarded as an improvement in the democratic process because the strength of preferences can be allowed to affect the outcome. Hayek, on the other hand, argues that the trade in vote support contributes to results that do not reflect the opinions of the majority. These views are clearly contradictory, yet the abstract nature of Hayek's analysis of the democratic process makes a comparison with conventional public finance difficult. However, Hayek's criticism of trade in vote support within representative assemblies is directed at the wide range of issues on which such practices are possible rather than at the principle of logrolling itself. The argument is derived from the distinction between votes on general rules of just conduct applicable to *all* citizens (i.e., *legislation* proper, such as private and criminal law) and votes on the distribution of payments and benefits among individuals and groups in society. The latter include all votes on taxes and transfer payments, allocation decisions with income redistribution implications, and "directives" that discriminate in favor of certain groups. The representative assemblies in all democracies do in fact practice logrolling on both types of issues. It is, therefore, possible that the distribution of benefits

[1] This work includes such classics as *The Road to Serfdom* (1944) and *The Constitution of Liberty* (1960).

[2] See, e.g., Tullock 1969.

on which a majority decide at the same time violates general rules on which the same majority agree. For example, a principle of taxation can be accepted while special benefits (e.g., tariffs and subsidies) with implicit positive or negative tax equivalents are passed by the same assembly, violating the agreed-on principle. It appears that the theory of logrolling is valid for votes on the distribution of benefits and payments *only* when no general agreed-on principle is applicable. Only pure collective goods are likely to satisfy this condition when a principle of taxation has been laid down, because individuals can be assigned tax equivalents for other publicly provided services and benefits.

The coercive powers of governments to benefit some groups at the expense of others, in addition to enforcing generally agreed-on principles, is the source of what Hayek calls "the miscarriage of the democratic ideal" (p. 98). These powers make it possible and advantageous to trade in specific benefits as well as to trade in benefits for support of general principles. Moreover, well-organized interests are likely to gain disproportionate weights in the democratic process—"there is no limit to the blackmail to which governments will be subject" when "no superior judiciary authority can prevent the legislature from granting privileges to particular groups" (p. 11).[3] The final outcome of the democratic process need not, then, correspond to anybody's opinion on what is right or to the will of the majority.[4] Hayek argues, in fact, that this concept *can* be stated only in terms of general rules and principles.[5] Behavioral incentives of the rules are lost and the will of the majority cannot be given contents when representatives decide *both* on general rules and on adjustments to the outcome of individuals' activities that are performed subject to these rules.

A Model Constitution

Hayek outlines a constitution designed both to make the outcome of the democratic process coincide with principles held by the majority and to minimize the degree of coercion in society.[6] The proposal is very simple: Distinguish between legislation proper and decisions on government spending, administration, and regulation by separating the two functions into two distinct assemblies elected by entirely different procedures. Make the government body subject to the general rules of conduct decided upon by the

[3] A suggestion like Friedman and Friedman's (1980) to legally prohibit the use of tariffs for the benefit of special-interest groups can obviously not get to the root of the problem of the democracy in Hayek's view as long as the government has the coercive power to use a multitude of alternative means of creating special benefits for the same groups.

[4] Note that Hayek regards the state of democracy in the United States as threatened as are the parliamentary democracies in Europe because the judiciary branch of the American government does generally not subject decisions of Congress to a test against general principles—though this may have been the intention of the Founding Fathers.

[5] This argument is developed in vol. 2 of *LLL*.

[6] Fritz Machlup has pointed out to me that one could argue that Hayek's proposal for constitutional reform contradicts his own views on the creation of human institutions. While Hayek argues that human institutions cannot be designed rationally and deliberately for the benefit of mankind, he also argues that they cannot develop spontaneously in accordance with people's preferences unless the conditions for such developments exist. The proposal constitutes a deliberate design for the purpose of creating such conditions.

legislative assembly. Hayek's model constitution also contains a general declaration of rights and an important definition of what qualifies as law—a general rule of just conduct. A constitutional court would be established to test the appropriateness of the legislature's decisions against such a definition.

The *legislative assembly* would be responsible for the body of criminal and private law, the principles of taxation, general regulations for safety and health, rules to secure competitive markets, corporate law, and the like. The coercive powers of governments would be limited to the enforcement of these general rules and principles.

The *government assembly* would decide on the use of material and human resources entrusted to the public sector. The size and the general purposes of expenditures would be limited only indirectly by the general rules of conduct set down by the legislative assembly and by people's willingness to pay taxes. The general principles of taxation decided upon in the legislative assembly would make citizens aware of their share in payments for specific services. This would prevent the current practice of disguising tax burdens to "make those who will ultimately have to bear it [the burden] as little aware of it as possible" (p. 127).

Would the proposed constitution achieve its purposes? A critical issue seems to be whether the legislative assembly can be prevented from instituting laws favoring large groups or wealthy interests. Hayek presents a number of suggestions in this regard. More important, however, is the definition of "law" and the role of the constitutional court in evaluating the constitutionality of legislation against this definition. Hayek characterizes (p. 109) a law as negative in the sense that it does not "aim at achieving concrete purposes," it must be applicable to an "indefinite number of unknown future instances," and it must exclude "all provisions intended or known to affect principally particular identifiable individuals or groups." This definition would apparently rule out legislation covering such issues as minimum wages and price controls. However, the boundary between measures to improve information spreading in the market and the protection of particular firms or groups is sometimes extremely vague, as in the case of legislation on health standards, consumer protection, and occupational safety. It is therefore easy to imagine situations in which the members of the constitutional court assume a critical role. The experience with the U.S. Supreme Court indicates that the existence of a constitutional court does not necessarily provide a complete safeguard against legislation that favors large groups in society. A simple addition to Hayek's proposal could help achieving his purpose—Wicksell's old suggestion (Wicksell 1896) of qualified majorities and relative unanimity. A principle that legislation must be passed by, for example, 90 percent of the assembly hardly seems objectionable since decisions on general rules must reflect the standards of most citizens to be respected and justify coercion in enforcement.

Where does Hayek's proposal lead? Debate on it cannot be carried out without reference to more specific results. Since the author provides only general and abstract clues, I will here attempt to illustrate some potential consequences by relating it to a few economic political issues.

The inflationary bias of current democracies has been a theme in Hayek's earlier writings and is clearly linked to the subject of this volume since inflation is one way for governments to disguise the true tax burden. The independence of the central bank from the government seems guaranteed under the proposed constitution because directives covering the central bank will have to take the form of general rules unrelated to specific political and economic circumstances. This provides a check on the monetization of gov-

ernment fiscal deficits.[7] Inflationary pressures could also be reduced under the proposed constitution because the government would be unable to bear the costs of unemployment and business failures. Most market interventions in the form of subsidies to labor or capital are likely to violate the general rules by which the government must abide. This could contribute to higher wage and price flexibility.[8] Furthermore, governments must consider its expenditures more carefully when the corresponding tax burden for each individual has been established by the legislature.[9]

A more controversial consequence of the proposed constitution would be that developments in production, consumption, and income distribution to which a majority object have to be accepted unless a general rule has in fact been violated. However, the outcome of economic activities may shed new light on a particular principle and lead to a change therein. The extremely far-reaching implications of this can be illustrated with the examples of immigration laws and equity-oriented policies.

Immigration restrictions appear to be unconstitutional, as do tariffs or import quotas, under Hayek's proposal. Such policy measures are clearly directed at identifiable individuals to achieve concrete purposes—economic benefits for national residents and firms. Opening up the Western industrialized economies to unlimited immigration would lead to a social transformation of enormous proportions. Most people would probably find the outcome unacceptable. Nevertheless, they would have to accept immigration as the "will of the majority" unless a general principle could be found to prevent immigration without thwarting other desirable outcomes, such as imports of competitive foreign goods.

Most policies aimed at securing certain levels of incomes for individuals or groups also have to be given up by governments that must abide by general principles. Discrimination is necessary to guarantee the outcome of economic activities pertaining to an individual. This is illustrated in volume 2 of *LLL* by a revealing and sharply satirical analysis of the contents of the United Nations *Universal Declaration of Human Rights*. This declaration contains positive rights (to outcomes) as well as traditional civil rights (to equal treatment before the law as defined by general rules of conduct). The two kinds of rights are simply contradictory. Either governments treat unequal people equally, abiding by generally accepted rules, or they treat unequal people unequally to secure certain outcomes. The only available policy instrument aimed at securing a certain income distribution would be the general tax structure. Thus, a negative income tax would have to be substituted for large parts of the social security system.

The examples above suffice to show that the constitutional proposal is far from politically feasible in the foreseeable future. There is probably no significant political group in any democracy that bases its political targets on

[7] Hayek earlier suggested the denationalization of the right to issue money (Hayek 1976*a*). Though a possibility, this suggestion is not part of the proposal here.

[8] "Social responsibility" must be the major check on wage and cost increases in an economy in which the government guarantees employment independent of the wage/cost structure (cf. Wihlborg 1978).

[9] Hayek's proposal can here be compared with the current discussion about a constitutional amendment prohibiting an unbalanced government budget. Just as tariffs cannot reduce the political pressures to discriminate in favor of a particular industry, a constitutional amendment to balance the budget does not reduce the political incentive to tax via inflation. Thus, the political process can be used, e.g., to manipulate the definition of the government budget—rendering the amendment meaningless.

general principles rather than on desired outcomes. Moreover, the proposal does not address the question of whether *any* outcome under a certain principle is acceptable. For example, is it not necessary to discriminate in order to secure survival of an individual who spends the negative income tax payments almost on receipt? Should hospital care be denied individuals who have chosen not to invest in medical insurance? Hayek has previously suggested ways of resolving these questions with a minimum of government coercion (Hayek 1960). It is not clear, however, that these solutions are commensurate with strict adherence to the constitutional principles outlined here. Despite such questions, Hayek's proposal deserves serious consideration as the starting point for further work by economists and social philosophers. A fundamental constitutional change may indeed be the only route to prevent further erosion of democratic values and individual liberty. Hayek has argued—convincingly, in my view—that the coercive powers given to governments, in the belief that progress can and should be planned and directed, are the cause of a self-generating process leading to the destruction of the decentralized market economies in which a high degree of individual liberty remains feasible.

Hayek's constitutional proposal implies quite simply that we give up the idea of steering and planning the future direction of social and economic activities. Instead, we should ask how we can set up a system of general principles under which mankind's knowledge can ensure progress for the maximum benefit of all. In Hayek's words: "To pretend to know the desirable direction of progress seems to me to be the extreme of hubris. . . . All we can do is to create favorable conditions for it [progress] and then hope for the best" (p. 169).

CLAS WIHLBORG

New York University

References

Friedman, Milton, and Friedman, Rose. *Free to Choose: A Personal Statement.* New York: Harcourt Brace Jovanovich, 1980.

Hayek, Friedrich A. *The Road to Serfdom.* Chicago: Univ. Chicago Press, 1944.

———. *The Constitution of Liberty.* Chicago: Univ. Chicago Press, 1960.

———. *Law, Legislation and Liberty: A New Statement of the Liberal Principles of Justice and Political Economy.* Vol. 1, *Rules and Order.* Chicago: Univ. Chicago Press, 1973.

———. *The Denationalization of Money: An Analysis of the Theory and Practice of Concurrent Currencies.* London: Inst. Econ. Affairs, 1976. (*a*)

———. *Law, Legislation and Liberty: A New Statement of the Liberal Principles of Justice and Political Economy.* Vol. 2, *The Mirage of Social Justice.* Chicago: Univ. Chicago Press, 1976. (*b*)

Tullock, Gordon. "Problems of Majority Voting." In *Readings in Welfare Economics*, edited by the American Economic Association. Homewood, Ill.: Irwin, 1969.

Wicksell, Knut. *Finanztheoretische Untersuchungen nebst Darstellung und Kritik des Steuerwesens Schwedens.* Jena: Fischer, 1896. English translation by Richard A. Musgrave and Alan T. Peacock, eds., *Classics in the Theory of Public Finance.* London: Macmillan, 1958.

Wihlborg, Clas. "Liberty and Labor Markets: A Reflection on the Swedish Experience." *ACES Bull.* 20 (Fall/Winter 1978): 85–100.

Studies in the Economics of Search. Edited by STEVEN A. LIPPMAN and JOHN J. MCCALL.
Amsterdam: North-Holland Publishing Co., 1979. Pp. 225. $41.50 (cloth).

As a sequel to their excellent survey of the job search literature, Lippman and McCall have organized an impressive collection of new papers that materially enrich the subject. The collection includes original contributions by nine young authors on topics that range from the estimation of the elasticity of unemployment duration with respect to unemployment compensation to a theoretical demonstration that searching for a bargain is the real world substitute for Walras's auctioneer. The individual essays are of uniformly high quality as are the editing and production of the volume.

According to the editors, each paper is motivated by a desire either to test the empirical implications of the so-called "standard" search model or to extend its structure for the purpose of deriving a richer menu of behavioral implications. Regarding these motives as noble, I have attempted to point out the sense in which and the extent to which each author has been led by them. Since the essays are related to the standard model in the inverse order of their presentation in the volume, my comments read from back to front.

The standard model in a labor-market context supposes that an individual worker samples wage rates sequentially from an urn of offers with a known frequency distribution at a rate of one per period. The worker wishes to maximize expected lifetime earnings net of the out-of-pocket and opportunity costs of search. As is well known, the optimal strategy is to accept the first offer in excess of some reservation wage. The reservation wage equates the cost of taking one more sample to the expected gain attributable to another sample. Under reasonable conditions, the reservation wage is independent of search duration to date

That a reduction in the cost of search yields an increase in the reservation wage is a straightforward implication Given the standard model, an increase in the reservation wage, in turn, implies an increase in the expected time required to find an acceptable wage and an increase in expected earnings. Because unemployment compensation must be forgone when an offer is accepted, it acts as a subsidy to search in the standard model

Kathleen Classen tests for these predicted effects by estimating the reduced-form relationship between both unemployment duration and post-unemployment earning and unemployment compensation. Both the weekly benefit received and the maximum duration of benefits are included as right-hand-side variables. The analysis yields a strong positive relationship between duration and both measures of unemployment compensation liberality but no statistically significant relationship between eventual earning and liberality. Although the second result can be explained away, it is also true that alternative explanations exist for a positive effect on duration. The paper has many interesting things to say, but as a test of search theory, the results are inconclusive.

One would prefer to test directly for the implied relationship between the reservation wage and the cost of search A major problem in doing so is the nonobservability of the reservation-wage rate. Nick Kiefer and George Neumann report on an attempt to get around this problem. Their method takes advantage of the structure of the standard model. In the paper, they show that data on accepted wage rates and realized unemployment durations can be used to estimate simultaneously the structural parameters of both the reservation-wage equation and the probability of finding acceptable employ-

ment function. Although their estimate of the unemployment benefit effect is positive in the first case and negative in the second as expected, the estimated effects of maximum benefit duration have the "wrong" signs. In spite of this ambiguity, the paper represents a valuable contribution to the methodology of estimating sequential, discrete choice models.

The standard model "explains" unemployment as a rational response to wage dispersion. Critics argue that this explanation is of little importance, because workers with long unemployment spells, those that account for most unemployment, do not turn down offers. They simply cannot find a job. By allowing unemployed workers to allocate time between search effort and nonmarket activity, Ken Burdett has extended the relevance of the search theoretic framework even to the case in which there is no wage dispersion. In his model, a reduction in job availability on unemployment is amplified by the rational allocation of a worker's time.

In the paper presented in the volume, Burdett shows that search intensity increases, and the reservation wage falls with unemployment duration, as the unemployment-benefit period is exhausted. This theoretical result suggests the existence of misspecification bias in both of the empirical papers discussed above. The allocation of time among search, work, and leisure by an employed worker is also investigated. This case is a natural synthesis of classical labor supply and search theory. The principal contribution of the paper is to incorporate search theoretic considerations into a more general dynamic theory of labor supply.

The papers by Jeffrey Hall, Steven Lippman, and John McCall and by John Danforth study the consequences of relaxing the assumption of risk neutrality. Presuming decreasing absolute risk aversion, the implied relationship between the reservation wage and nonhuman wealth is positive. Hence, the papers represent a rigorous derivation of the long-standing contention that the running down of financial assets explains the observed negative association between the reservation wage and duration of unemployment. As a technical matter, Hall et al. show that the reservation-wage property need not hold when past offers can be recalled. Nevertheless, the probability per period of finding acceptable employment decreases with wealth in this case as well. Finally, Danforth's formulation represents a synthesis of the theory of household wealth accumulation and the theory of wage search. As Danforth points out, the theory implies a process of increasing wealth dispersion over time if the assumption of decreasing absolute risk aversion is empirically valid. Hence, the paper is also a contribution to the theory of income distribution.

The standard search model characterizes the optimal strategy of a price-taking agent seeking a bargain. The market effect of such behavior is a reallocation of goods and services toward demanders who value them most highly. One might expect that such a process would eventually equalize marginal valuations, that is, converge to a competitive equilibrium. Joseph Wharton formalizes this idea in the context of a market for homogeneous labor services composed of workers who search on the job for a higher wage and employers who adjust wage offers in response to search-induced changes in employment. The process is modeled as a Markov chain. When the cost of search is "small," a competitive allocation of workers across firms is the only absorbing state. Although the analysis is insightful and the approach developed is amenable to further development, the employer's wage-setting behavior is not shown to be rational in any sense. Other work on this problem suggests that strategic behavior on the part of the price-setting agents can thwart convergence to competitive equilibrium.

Search-induced convergence to a single price raises the following paradox. How can the standard search model represent an important contribution to positive economics if the market effect of the behavior explained eliminates the price dispersion required to observe the behavior? Roger Kormendi provides a resolution of this paradox by formulating a market model in a pure-exchange context that exhibits price dispersion in equilibrium. One of the original features of his model is that every agent plays a role in the price formation process. The demonstration that mutually consistent (Nash) price-setting strategies exist which define a market equilibrium is a major contribution of the paper. Although a detailed discussion of the kinds of conditions one needs for equilibrium price dispersion in Kormendi's model is beyond the scope of this review, it is important to note that nonstorability of the two commodities traded is implicit. If this assumption were relaxed, Kormendi's equilibrium is only "temporary" in the sense that, once allocations have changed in response to trade, the agents have incentives to play the game again. One suspects that the repeated game might well converge eventually to a competitive single price equilibrium.

It is curious that the matching of job and worker is not explicitly recognized in the standard formulation of the individual's problem of finding acceptable employment. Louis Wilde contributes to this deficiency by distinguishing between two job characteristics: the wage, observable on initial contact, and an attribute that must be experienced on the job. The analysis characterizes an individual worker's acceptance strategy assuming knowledge of the joint probability distribution over jobs. The result is a formalization of the "job-shopping" hypothesis typically offered as an explanation for high quit rates among the young. A useful extension of this model would also recognize the employer's problem of learning about a worker's productivity.

In the final paper, Charles Stuart applies the standard search model to an entirely different subject, the spatial clustering of sellers of both the same and different commodities. Specifically, he shows that market centers are more attractive to buyers than spatially isolated sellers, because the cost of search per unit is smaller and the perceived heterogeneity of the goods offered is larger. These sources of agglomeration economies offset the effects of transportation costs and competition that would otherwise induce dispersion in the location of sales outlets. Although there are many insightful observations on the subject, the author does not attempt to derive an equilibrium solution to the location problem in which all of these forces balance. This would seem to be a fruitful area for further research.

In sum, there is much of value in the volume both to students of labor economics and to those interested in applying search theory in other areas of economic inquiry.

DALE T. MORTENSEN

Northwestern University

A Theory of Income Distribution. By HAROLD LYDALL.
New York: Oxford University Press, 1979. Pp. vii+326. $32.50.

Professor Lydall, of South African birth, received his higher education at Oxford. He has taught primarily in Britain and Australia, worked in international agencies in Geneva, and also visited extensively in the United States.

His methodology recalls the Burns and Mitchell pragmatic tradition of quantitative institutionalism. That is to say, Lydall begins by searching out statistical regularities in his data and then seeks one or more theories to explain them, preferably on more than an ad hoc basis. He looks with some suspicion on the opposite procedure of deducing a general theory first and then seeking out data to test it.

Although the present volume is indeed more nearly theoretical and less exclusively statistical than Lydall's well-known *Structure of Earnings* (1968), its title may seem misleading to the conventional economist. I myself expected a new "high theory" (Shackle's phrase), or at least a tasteful ragout of existing high theories. But after seven "background" chapters of a doctrinal-historical nature (which explain why the author feels it desirable to strike out along new paths), Lydall's independent contribution retreats more than half way to empiricism.

Once inoculated against anticipations of high theory—wage funds, iron laws, marginal productivities, widows' cruses, degrees of monopoly—the reader is impressed by Lydall's statistical generalizations, both in themselves and as challenges for integration into the rival high theories of our rival conventional wisdoms.

Lydall finds, for example, the following relation between value added per worker, which he calls v, and the average size of the employing firm as measured by employment, L: $v = BL^{\beta-1}$ ($B > 0$, $\beta > 1$). A similar relation holds for the average wage, w: $w = AL^{\alpha-1}$ ($A > 0$, $1 < \alpha < \beta$), and for one-man firms (OMF), $A = B$. The size distribution of firms follows the Pareto law, with N as the number of firms: $N = RL^{-\rho}$ ($R > 0, \rho > 1$). (In these three equations, both capital and Greek letters are statistical parameters.)

An ingenious algebraic development (pp. 297, 298) combines these results into a theoretical labor share S_u of $(\rho - \beta)/(\rho - \alpha)$, derived under conditions Lydall calls EPM (entry and product market) competition. In EPM, diminishing returns to scale need never hold as in conventional pure competition, but freedom of entry for tiny interstitial OMF is an adequate check on the abuses of monopoly and oligopoly—when coupled, I should myself add, with plausible threats of regulation or socialization. (Lydall sees this EPM-OMF framework as among his major contributions.) The EPM competition is atomistic but imperfect. It permits, in particular, imperfections in knowledge, unutilized economies of scale, and heterogeneous capital. It is consistent even in the long run, Lydall believes, with such phenomena as markup pricing, production under increasing returns, and persistent differentials in rates of return on investments of different sorts. As to the importance of these modifications, my reaction is wait and see, influenced more by the post-Chamberlin history of the "large group case" of monopolistic competition than by any analytical flaw in EPM-OMF or its applications.

Two further innovations in Lydall's approach and results relate to the dynamics of distribution and development. They explore the Kuznets conjecture that, in Lydall's words (pp. 129, 130), "the degree of inequality is fairly low in the poorest countries, increases with per capita income up to a certain level, and thereafter follows a continuous downward trend." In a cross-country study, a Lydall student (T. Das) derived an empirical estimate of the determinants of the Gini concentration ratio, G, with dependent variables X (income per capita), its reciprocal, and four D_i or "dummy" variables:[1] $G = 0.784 - 0.050 \log X - 12.184X^{-1} + 0.051D_1 + 0.124D_2 - 0.033D_3 - 0.007D_4$

[1] I interpret D_1 as the proportion of income recipients (to total population), D_2 as the proportion of economically active persons, D_3 as the proportion in the rural (or

Lydall provides later (pp. 219, 220) his own explanation of the Kuznets phenomenon. He has divided each economy under study into a traditional sector, 1, and a modern sector, 2. The average income, Y_i, of each sector is related logarithmically to overall average income, $\bar{Y}$: $\log Y_1 = 1.457 + 0.725 \log \bar{Y}$, $\log Y_2 = 3.86 + 0.55 \log \bar{Y}$; we also have $\bar{Y} = (1 - p)Y_1 + pY_2$, with the proportion p rising as development progresses. Lydall assumes income distributed lognormally in each sector with a uniform logarithmic standard deviation. From these results and assumptions he derives by simulation (his table 12.1) numerical results which not only conform to the Kuznets hypothesis but go beyond it in one important respect. The share of the lowest 20 percent of the population eventually turns up as development proceeds, which it did not appear to do in Kuznets's original data.

The most noteworthy feature of Lydall's discussion of personal income distribution (chaps. 12–14) is its conscious endeavor to redress the balance between competitors already in the field. As economists know, the human-capital theories of T. W. Schultz, Gary Becker, and others have risen to dominance over the past quarter century. Lydall represents a reasoned and good-natured rebuttal in the ongoing debate. At the same time, he avoids what I should call the major error of Jan Tinbergen's 1975 distribution monograph, which is to ignore nonlabor incomes entirely as atavistic residuals destined to depart the scene in the near future.

Postwar distribution theory has been marked by a number of efforts to bridge the gaps and construct at least pontoon bridges between deductive functional theories and empirical personal ones. These efforts have ranged in America alone from the "human-capital" approach mentioned above to "job-ladder" and "labor-market segmentation" approaches generally more critical of the status quo. Lydall's study is a major and possibly a highly influential addition to this literature, although Lydall's own summary of his innovative conclusions (chap. 15, esp. pp. 285–91) is occasionally less than generous to his numerous predecessors. If Lydall's book is intended as more than a contribution, but as the late twentieth-century equivalent of John Bates Clark's late nineteenth-century magnum opus (as that summary chapter hints), I fear that it falls somewhat short.

MARTIN BRONFENBRENNER

Duke University

---

agricultural) sector, and D_4 as the proportion in the urban (or nonagricultural) sector (The statistical significance of D_3 and D_4 appears dubious in Lydall's text.) The R^2 value of 0.38, while low, is significantly greater than zero.

LIFE CHANCES

Approaches to Social and Political Theory

Ralf Dahrendorf

Dahrendorf considers the fundamental questions of the meaning of history and the viability of the notion of social progress through the concept of "life chances"—human options limited by social ties, obligations, and expectations.

Cloth $15.00 Paper $5.95 192 pages

LAW, LEGISLATION, AND LIBERTY
Volume 3

The Political Order of a Free People

F. A. Hayek

". . . Hayek offers an acute defense of the contemporary relevance of classical economic principles against the criticisms of other economists."—*American Political Science Review*

Cloth $14.00 Paper $7.50 260 pages

THE POLITICAL SYSTEM

An Inquiry into the State of Political Science

Second Edition

David Easton

"David Easton has brought together in this book the results of sustained and vigorous thinking on his part about the dilemmas of political science, and has attempted to sketch the requirements of more coherent and cumulative work for the future."—Oliver Garceau, *Political Science Quarterly*

Paper $12.50 408 pages

From Midway Reprints—

INDIVIDUALISM AND ECONOMIC ORDER

Friedrich A. Hayek

Paper $13.00 280 pages

THE UNIVERSITY OF CHICAGO PRESS

5801 South Ellis Avenue Chicago, Illinois 60637

The Role of Market Forces in Assuring Contractual Performance

Benjamin Klein

University of California, Los Angeles

Keith B. Leffler

University of Washington

The conditions under which transactors can use the market (repeat-purchase) mechanism of contract enforcement are examined. Increased price is shown to be a means of assuring contractual performance. A necessary and sufficient condition for performance is the existence of price sufficiently above salvageable production costs so that the nonperforming firm loses a discounted stream of rents on future sales which is greater than the wealth increase from nonperformance. This will generally imply a market price greater than the perfectly competitive price and rationalize investments in firm-specific assets. Advertising investments thereby become a positive indicator of likely performance.

I. Introduction

An implicit assumption of the economic paradigm of market exchange is the presence of a government to define property rights and

We thank Armen Alchian, Thomas Borcherding, Harold Demsetz, James Ferguson, Jack Hirshleifer, Matt Lindsay, Roy Kenney, John Long, Ian Macneil, Kevin Murphy, Phillip Nelson, Joseph Ostroy, Peter Pashigian, Sam Peltzman, George Priest, John Riley, Jonathan Skinner, George Stigler, Earl Thompson, and participants at seminars at UCLA and the University of Chicago during 1977–78 for helpful suggestions and comments. Jonathan Skinner also provided valuable research assistance. The Foundation for Research in Economics and Education and the University of Chicago Law School Law and Economics Program provided Klein with research support, and the Center for Research in Government Policy and Business, University of Rochester, provided Leffler with research support.

[*Journal of Political Economy*, 1981, vol. 89, no. 4]
© 1981 by The University of Chicago 0022-3808/81/8904-0004$01.50

enforce contracts. An important element of the legal-philosophical tradition upon which the economic model is built is that without some third-party enforcer to sanction stealing and reneging, market exchange would be impossible.[1] But economists also have long considered "reputations" and brand names to be private devices which provide incentives that assure contract performance in the absence of any third-party enforcer (Hayek 1948, p. 97; Marshall 1949, vol. 4, p. xi). This private-contract enforcement mechanism relies upon the value to the firm of repeat sales to satisfied customers as a means of preventing nonperformance. However, it is possible that economic agents with well-known brand names and reputations for honoring contracts may find it wealth maximizing to break such potentially long-term exchange relationships and obtain a temporary increase in profit. In particular, the determinants of the efficacy of this market method of contract performance and therefore the conditions under which we are likely to observe its use remain unspecified.

This paper examines the nongovernmental repeat-purchase contract-enforcement mechanism. To isolate this force, we assume throughout our analysis that contracts are not enforceable by the government or any other third party. Transactors are assumed to rely solely on the threat of termination of the business relationship for enforcement of contractual promises.[2] This assumption is most realistic for contractual terms concerning difficult-to-measure product characteristics such as the "taste" of a hamburger. However, even when the aspects of a contract are less complicated and subjective and therefore performance more easily measurable by a third party such as a judge, specification, litigation, and other contract-enforcement costs may be substantial. Therefore, explicit guarantees to replace or repair defective goods (warranties) are not costless ways to assure contract performance. Market arrangements such as the value of lost repeat purchases which motivate transactors to honor their promises may be the cheapest method of guaranteeing the guarantee.

While our approach is general in the sense that the value of future exchanges can motivate fulfillment of all types of contractual prom-

[1] Hobbes ([1651] 1955, pp. 89–90) maintains that ". . . he that performeth first, has no assurance the other will perform after; because the bonds of words are too weak to *bridle men's ambition, avarice, anger, and other Passions, without the fear of some* coercive Power; which in the condition of here Nature, where all men are equal, and judges of the justness of their own fears cannot possibly be supposed."

[2] This assumption is consistent with the pioneering work of Macaulay (1963), where reliance on formal contracts and the threat of explicit legal sanctions was found to be an extremely rare element of interfirm relationships. Macaulay provides some sketchy evidence that business firms prevent nonfulfillment of contracts by the use of effective nonlegal sanctions consisting primarily of the loss of future business. This "relational" nature of contracts has been recently emphasized by Macneil (1974), and also by Goldberg (1976) and Williamson (1979).

ises, we focus in this paper on contracts between producers and consumers regarding product quality. In order for a repeat-sale enforcement mechanism to operate, we assume that the identity of firms is known by consumers[3] and that the government enforces property rights to the extent that consumers voluntarily choose whom to deal with and must pay for the goods they receive.[4] In addition, managers of firms are assumed to be wealth maximizing and to place no value on honesty per se.

In Section II, the conditions are outlined under which firms will either honor their commitments to supply a high level of quality or choose to supply a quality lower than promised. In order to emphasize the ability of markets to guarantee quality in the absence of any government enforcement mechanism, a simple model is presented which assumes that consumers costlessly communicate among one another. Therefore, if a firm cheats and supplies to any individual a quality of product less than contracted for, all consumers in the market learn this and all future sales are lost. A major result of our analysis is that even such perfect interconsumer communication conditions are not sufficient to assure high quality supply. Cheating will be prevented and high quality products will be supplied only if firms are earning a continual stream of rental income that will be lost if low quality output is deceptively produced. The present discounted value of this rental stream must be greater than the one-time wealth increase obtained from low quality production.

This condition for the "notorious firm" repeat-purchase mechanism to assure high quality supply is not generally fulfilled by the usual free-entry, perfectly competitive equilibrium conditions of price equal to marginal and average cost. It becomes necessary to distinguish between production costs that are "sunk" firm-specific assets and those production costs that are salvageable (i.e., recoverable) in uses outside the firm. Our analysis implies that firms will not cheat on

[3] Nonidentification of firm output leads to quality depreciation via a standard externality argument; i.e., supply by a particular firm of lower than anticipated quality imposes a cost through the loss of future sales not solely on that firm but on all firms in the industry (see Akerlof 1970; Klein 1974)

[4] For simplicity, we assume that "theft," as opposed to nonfulfillment of contract, is not possible. While "fraud," in the sense of one party to the transaction intentionally supplying less than contracted for, is analytically similar to "theft," we draw a distinction along this continuum by assuming that the government only permits "voluntary" transactions in the sense that transactors choose whom to trade with. Therefore, while consumers cannot "steal" goods, they can, in principle, pay for the goods they receive with checks that bounce, and while firms cannot rob consumers, they can, in principle, supply goods of lower than promised quality. Although we recognize the great difficulty in practice of separating the underlying government enforcement mechanisms, e.g., property law, from the private promise-enforcing mechanisms we are attempting to analyze, this distinction between theft and fraud is analytically unambiguous.

promises to sell high quality output only if price is sufficiently above salvageable production costs. While the perfectly competitive price may imply such a margin above salvageable costs, this will not necessarily be the case. The fundamental theoretical result of this paper is that market prices above the competitive price and the presence of nonsalvageable capital are means of enforcing quality promises.[5]

In Section III our theoretical model of quality-guaranteeing price premiums above salvageable costs is extended to examine how the capital value of these price-premium payments can be dissipated in a free-entry equilibrium. The quality-guaranteeing nature of nonsalvageable, firm-specific capital investments is developed. Alternative techniques of minimizing the cost to consumers of obtaining an assured high quality are investigated. We also explore market responses to consumer uncertainty about quality-assuring premium levels. Advertising and other production and distribution investments in "conspicuous" assets are examined as competitive responses to simultaneous quality and production-cost uncertainties. Finally, a summary of the analysis and some concluding remarks are presented in Section IV.

II. Price Premiums and Quality Assurance

Assume initially that consumers costlessly know all market prices and production technologies but not the qualities of goods offered for sale. For simplicity, the good being considered is assumed to be characterized by a single objective quality measure, q, where quality refers to the level of some desirable characteristic contained in the good. Examples are the quietness of appliance motors, the wrinkle-free or colorfast properties of clothing, or the gasoline mileage of an automobile. We also assume that the economy consists of consumers who consider buying a product x each period, where the length of a period is defined by the life (repurchase period) of product x, and who are assumed to costlessly communicate quality information among one another. Therefore, if a particular firm supplies less-

[5] The notion that an increased price can serve as a means of assuring high quality supply by giving the firm a rental stream that will be lost if future sales are not made is not new. Adam Smith ([1776] 1937, p. 105) suggested this force more than 200 years ago when he noted that "the wages of labour vary according to the small or great trust which must be reposed in the workman. The wages of goldsmiths and jewellers are everywhere superior to those of many other workmen, not only of equal, but of much superior ingenuity; on account of the precious metals with which they are intrusted. We trust our health to the physician; our fortune and sometimes our life and reputation to the lawyer and attorney. Such confidence could not safely be reposed in people of a very mean or low condition." Similar competitive mechanisms recently have been analyzed by Becker and Stigler (1974) and Klein (1974).

than-contracted-for quality to one consumer, the next period all consumers are assumed to know. In addition, this information is assumed not to depreciate over time.[6]

Identical technology is assumed to be available to all entrepreneurs. Hence, there are many potential firms with identical total cost functions, $C = c(x,q) + F(q)$, where F is fixed (invariant to rate) costs. Higher quality and larger quantities require higher production costs, $F_q > 0$, $c_q > 0$, $c_x > 0$, and marginal cost is assumed to increase with quality, $c_{xq} > 0$. Fixed costs are assumed initially to be expenditures made explicitly each period rather than capital costs allocated to the current period. For example, they may include a payment on a short-term (one-period) rental agreement for a machine but not the current forgone interest on a purchased machine or the current period's payment on a long-term rental agreement—both of which imply long-term and hence capital commitments.

We therefore are explicitly distinguishing between "fixed" costs in the sense employed here of constant (invariant to output) current costs and "sunk" (nonsalvageable) capital costs. The usual textbook proposition that a firm will not shut down production as long as price is greater than average variable cost blurs this distinction and implicitly assumes that all fixed costs are also sunk capital costs. Our assumption of the complete absence of any long-term commitments is analytically equivalent to perfect salvageability of all capital assets. If all long-term production-factor commitments were costlessly reversible, that is, all real and financial assets such as the machine or the long-term machine rental contract could be costlessly resold and hence perfectly salvageable, there also would not be any capital costs. Only the nonsalvageable part of any long-term commitment should be considered a current sunk capital cost.

If buyers are costlessly informed about quality, the competitive price schedule, P_c, for alternative quality levels is given by the minimum average production costs for each level of quality and is designated by $P_c = P_c(q)$. This is represented in figure 1 for two alternative quality levels, q_h and q_{min}, by the prices P_1 and P_0. Suppose, however, that the quality of product x cannot be determined costlessly before purchase. For simplicity, assume prepurchase inspection reveals only whether quality is below some minimum level, q_{min}, and that

[6] If we modify the assumptions of our model to make interconsumer communication less than perfect and allow inflows of new ignorant consumers over time and permit individuals to forget, the potential short-run cheating gain by firms would be increased. Therefore, the quality-assuring price premium would be higher than we derive below. In this case increased firm size, by making it more likely that the individuals one is sharing product-quality information with (e.g., family and friends) have purchased from the same firm, lowers the potential short-run cheating gain by essentially reducing the repurchase period.

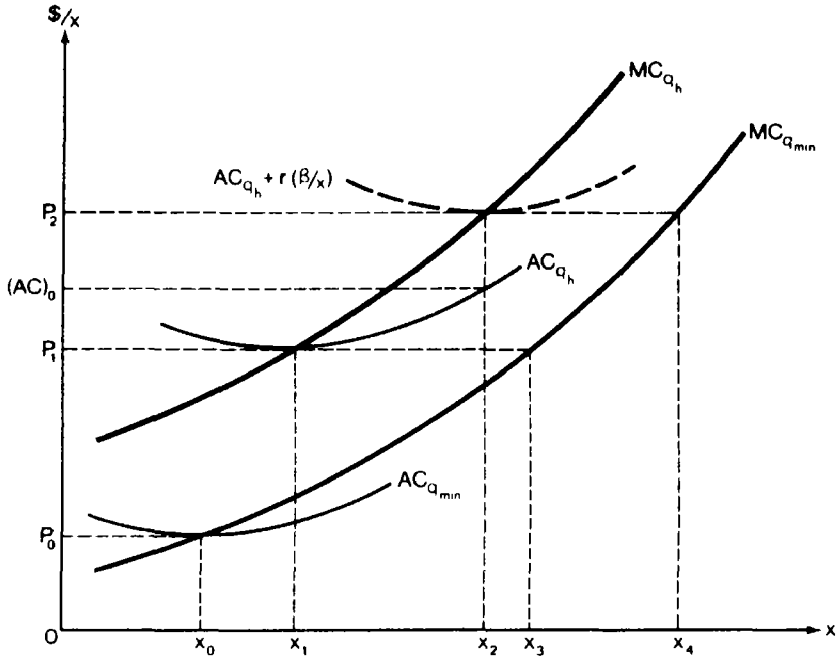


FIG. 1.—Pricing and production of alternative quality levels

the costs are prohibitive of determining whether quality is above q_{min} prior to purchase.[7] Obviously, whenever the market price that consumers will pay for asserted high quality exceeds the cost of producing minimum quality output, the firm can increase its initial period profits by producing the minimum quality output and deceptively selling it as a higher quality product.

If producers are to have an incentive to produce high quality products (in the absence of governmentally enforceable contracts), consumers must somehow reward high quality production and punish low quality production. We assume in this competitive framework that consumers will purchase from particular sellers randomly chosen from the group of homogeneous sellers over which consumer information is transmitted. If a consumer receives a product of a quality at least as high as implicitly contracted for, he will continue to purchase randomly from this group of sellers. On the other hand, if quality is less than contracted for, all consumers cease to purchase from the particular sampled "cheating" firm.

[7] The quality of the good beyond the minimum level is therefore what Nelson (1970) has labeled as an "experience" characteristic. Making the minimum quality level endogenous does not substantially change the following analysis.

Consider now an initial "competitive" equilibrium in which a single firm contemplates selling a quality below that expected by customers. Given the competitive market price for some high quality, $P_c(q_h) \equiv P_1$ in figure 1, this particular firm could increase its initial period quasi rents by producing minimum quality and selling it at the high quality price. However, since buyers are assumed to communicate fully with one another, all future customers of high quality output, that is, sales at prices greater than $P_c(q_{\min}) \equiv P_0$ in figure 1, are lost. That is, a firm that cheats will become known as a "notorious" cheater, and consumers will not purchase from the firm any product the quality of which cannot be determined prepurchase.[8]

Whether sales of high or minimum quality will maximize the firm's wealth depends on whether the capital value of future quasi rents from continued high quality production exceeds the differential initial period quasi rents from quality depreciation. In terms of figure 1, at the perfectly competitive price for high quality output, P_1, price is equal to the average costs of high quality production. Therefore, the quasi rents from continued high quality production are zero. If, alternatively, the firm were to deceptively produce minimum quality output, as a price taker it would expand its production to x_3 (where $P_1 = MC_{q_{\min}}$) and receive a one-period quasi rent, the present value of which is equal to:

$$W_1 = \frac{1}{1+r}\left\{(P_1 - P_0)x_3 - \int_{x_0}^{x_3}[MC_{q_{\min}}(x) - P_0]dx\right\}. \qquad (1)$$

Therefore, at the perfectly competitive price for any quality above $q_{\min}$ firms will always cheat consumers and supply $q_{\min}$.

Faced with this possibility, consumers would recognize that regardless of producers' promises they will not obtain the higher quality product. Therefore, consumers would be willing to pay only the costless information price of the minimum quality product whose quality they can verify prepurchase, P_0. Because of such rational consumer anticipations, firms will not be able to cheat, but desired high quality output will not be supplied.

There may, however, be a price higher than the perfectly competitive price of high quality output, P_1, that if it were the equilibrium market price would (*a*) motivate honest production of the high quality good and (*b*) not completely dissipate the consumers' surplus from purchase of higher quality. Consider a price such as P_2 in figure 1. A

[8] A terminated firm cannot begin business in this industry under a new name. However, the highest valued alternative use of the entrepreneurial skills is included in salvageable fixed production costs. The firm considered here is assumed to face the same opportunities elsewhere as the firms that are honest in production of x. Therefore, the cheating firm can elect to enter a new industry.

firm supplying high quality output will now expand its production to x_2. The price premium $\tilde{P}$, defined as the increase in the price above minimum average cost of high quality, provides firms supplying high quality with a perpetual stream of quasi rents the present value of which (assuming unchanging cost and demand conditions over time) is equal to:

$$W_2 = \frac{1}{r}\left\{\tilde{P}x_2 - \int_{x_1}^{x_2}[MC_h(x) - P_1]dx\right\}. \tag{2}$$

The price premium also increases the gains to a firm from supplying minimum quality at the high price. A firm that chooses to cheat will now expand its output (in terms of fig. 1 to x_4) and earn the extra premium on all units sold.[9] Therefore, the capital value of the quasi rents from supplying quality less than promised is:

$$W_3 = \frac{1}{1+r}\left\{[\tilde{P} + (P_1 - P_0)]x_4 - \int_{x_0}^{x_4}[MC_{q_{\min}}(x) - P_0]dx\right\}. \tag{3}$$

A firm will honor its implicit quality contract as long as the difference between the capital values of the noncheating and cheating strategies, $W_2 - W_3$, is positive. Consider the quasi-rent flow of the cheating and noncheating alternatives, that is, the terms in braces in our expressions (2) and (3). Define QR_2 equal to rW_2 and QR_3 equal to $(1 + r)W_3$. A firm will then elect not to cheat if and only if:

$$\frac{QR_3}{QR_2} \leqslant \frac{(1+r)}{r}. \tag{4}$$

Therefore, there will be a price premium that motivates firms to honestly produce high quality as long as:

$$\lim_{\tilde{P}\to\infty}\left(\frac{QR_3}{QR_2}\right) \leqslant \frac{(1+r)}{r}. \tag{5}$$

Using L'Hospital's rule, equation (5) will be satisfied as long as

$$\frac{1}{r} > \frac{(x_4 - x_2)}{x_2} \tag{6}$$

for all $P > P_3$, where P_3 is some finite price. Intuitively, as the price increases it is only the increase in quasi rents on the additional units of minimum quality output that favors the deceptive strategy. Equation (6) insures that price increases beyond some level increase W_2 more than W_3 such that eventually W_2 is greater than W_3.

[9] Note that although x_2 may be greater or less than x_3, depending on the price premium chosen, given upward-sloping supply functions and the condition that $MC_{q_h}(q) > MC_{q_{\min}}(q)$ for all q, it must be the case that $x_4 > x_2 > x_1$.

The condition specified in equation (6) is quite reasonable. It will be satisfied as long as a cheating firm does not accompany cheating with very large output increases. If, for example, the real interest rate were .05 we require only that the output increase by a cheating firm not be more than 20 times the total output that would be produced if the firm were not cheating. Hence, under very general cost conditions a price premium will exist that motivates competitive firms to honor high quality promises because the value of satisfied customers exceeds the cost savings of cheating them.[10]

While we cannot state broad necessary conditions for the form the cost function must take to imply the existence of a quality-assuring price, "reasonable" sufficient conditions can be stated. In particular, all cases of vertically parallel marginal cost curves, as illustrated in figure 1, where quality is produced by a fixed input not subject to decreasing returns to scale (such as the use of a better motor) and where the second derivative of marginal cost is greater than or equal to zero imply the existence of a quality-assuring price. The Appendix contains simulation results under the more unrealistic assumption of isoelastic marginal cost functions. These simulations indicate the exceptional nature of the conditions when equation (6) is not satisfied. When a quality-assuring price does not exist, the cost functions are generally such that at reasonable premiums cheating output would be expanded by very large factors (often factors of many thousands). Since marginal cost functions for most products can be expected to become quite steep if not vertical at output expansions of much less than these factors, a quality-assuring price premium can generally be expected to exist.

Throughout the remainder of the paper we assume the existence of a quality-assuring price. For given cost and demand conditions, the minimum quality-assuring price will depend upon the level of quality

[10] The potential function of price premiums as quality guarantors is also applicable to markets in which firms face downward-sloping demands. In this case, the inability of firms to increase sales without reductions in price limits the gains available from deceptive minimum quality production as price increases. The existence of a price sufficient to guarantee quality now depends on the elasticity of demand in addition to the cost savings from quality reductions at various quantities. In addition, when price-searching firms do not have stable future demands, consumer knowledge of cost and current demand conditions is not sufficient to estimate the quality-assuring price. The anticipated future demand vis-à-vis current demand is also relevant. For example, where consumers expect a growing demand for the output of a firm that continues to produce high quality output, the rate of quasi-rent flow from high quality (or future deceptive minimum quality) production increases over time. As compared to a firm with the same initial but constant demand, the growing firm will receive a larger capital value return at any price from high quality production in the initial period. Firms facing expected demand growth will therefore require smaller quality-assuring price premiums. See Klein, McLaughlin, and Murphy (1980) for an analysis of the less than perfectly elastic firm demand case.

considered and is denoted by $P^* = P^*(q, q_{min}, r)$. Our analysis implies that the quality-assuring price will increase as quality increases, as minimum quality decreases (for all q greater than q_{min}), and as the interest rate increases. These conditions are consistent with the familiar recognition that, given a particular quality level, quality-cheating problems are less severe the higher the level of quality that can be detected prepurchase and the shorter the period of repurchase.[11]

Intuitively, the quality-assuring price treats the potential value of not producing minimum quality as an explicit opportunity cost to the firm of higher quality production. Hence the quality-assuring price must not only compensate the firm for the increased average production costs incurred when quality above that detectable prior to purchase is produced, but must also yield a normal rate of return on the forgone gains from exploiting consumer ignorance. This price "premium" stream can be thought of as "protection money" paid by consumers to induce contract performance. Although the present discounted value of this stream equals the value of the short-run gain the firm can obtain by cheating, consumers are not indifferent between paying the "premium" over time or permitting the firm to cheat. The price "premium" is a payment for high quality in the face of prepurchase quality-determination costs. The relevant consumer choice is between demanding minimum quality output at a perfectly competitive (costless information) price or paying a competitive price

[11] We can complicate our model by dropping the assumption that nondeceiving firms are anticipated to produce forever. If firms have a finite life, and the last period of production is known by both firms and consumers, there will be no premium sufficient to guarantee quality. No matter how high the premium paid by consumers for a high quality good in the last period, firms will supply "deceptive" minimum quality because there are no future sales to lose. Consumers aware of the last period will therefore demand only the minimum quality in that period. But then the next to the last period becomes the last period in the sense that, independent of the price premium, firm wealth is maximized by supplying minimum quality and going out of business. Consumers will then only pay for minimum quality output in the next to last period, and so on. High quality will never be produced. However, the necessary unraveling of the premium solution to assure high quality requires prior consumer knowledge of a date beyond which the firm will not produce. If consumers merely know that firms have finite lives but cannot with certainty specify a date beyond which a particular firm will not exist, price premiums may assure quality. While consumers are aware that some transactions will be with firms in their last period and hence cheating will occur, the expected gain from purchasing high promised quality can be positive. Our price premium–repeat business quality enforcement mechanism is analytically equivalent in form to the "super-game" solutions to the prisoner's dilemma problem developed in the game-theory literature. A general result of that analysis is that a cooperative solution can exist if one assumes either an infinitely long super game (as we have assumed in our model), or a super game of finite length but with transactors who have sufficient uncertainty regarding the period when the super game will end (see, e.g., Luce and Raiffa [1957], pp. 97–102, or, for a more recent solution to the problem that is similar in emphasis to our approach, Telser [1980]).

"premium," which is both necessary and sufficient, for higher quality output.[12]

There is a possibility that the required quality-guaranteeing price premium may exceed the increased consumer surplus of purchasing higher quality rather than the *minimum* quality product. If consumers can easily substitute increased quantity of the low quality product for increased quality, then the value of guaranteed high quality will be relatively low. Therefore, although a quality-guaranteeing price exists, a higher than minimum quality product may not be produced. For those goods where the substitution possibilities between quality and quantity are lower (e.g., drugs), consumer demand for confidence will be relatively high and the high quality guarantee worth the price premium. We assume throughout that we are dealing with products where some demand exists for the high quality good in the range of prices considered.

III. Competitive Market Equilibrium: Firm-specific Capital Investments

Our analysis has focused on the case where costless information (perfectly competitive) prices do not imply sufficient firm-specific rents to motivate high quality production. A price premium was therefore necessary to induce high quality supply. Thus, if price assures quality, the firms producing quality greater than q_{min} appear to earn positive economic profits. However, this cannot describe a full market equilibrium. When the price is high enough to assure a particular high level of quality, additional firms have an incentive to enter the industry. But if additional firms enter, the summation of the individual firms' outputs will exceed the quantity demanded at the quality-assuring price. Yet this output surplus cannot result in price reductions since the quality-assuring price is, in effect, a minimum price constraint "enforced" by rational consumers. All price-taking firms supplying a particular promised quality q above the minimum face a horizontal demand curve at $P^* = P^*(q)$, which is of an unusual nature in that prices above or below P^* result in zero sales. Consumers know that any price below P^* for its associated quality results in the

[12] As opposed to the Darby and Karni (1973) analysis, this particular model implies an equilibrium quantity of "fraud" equal to zero, where fraud is the difference between anticipated and actual quality. Given the symmetrical information assumptions regarding cost functions, parties to a contract know when and by how much a contract will be broken. An unanticipated broken quality contract is therefore not possible. The implicit economic (as opposed to common usage) concept of "contract" refers to anticipated outcomes and not to verbal promises or written agreements; thus there will be no broken quality "contracts."

supply of q_{min}. They therefore will not purchase from a firm promising that quality at a price lower than P^*.

A. *Brand Name Capital Investments*

Competition to dissipate the economic profits being earned by existing firms must therefore occur in nonprice dimensions. However, the zero-profit equilibrium is consistent with only a very particular form of profit-absorbing nonprice competition. The competition involves *firm-specific capital* expenditures. This firm-specific capital competition motivates firms to purchase assets with (nonsalvageable) costs equal to the capital value of the premium rental stream earned when high quality is supplied at the quality-assuring price. That is, if $P^*(q)$ is not to increase, the investment leading to zero profits must be highly firm specific and depreciate to zero if the firm cheats and supplies q_{min} rather than the anticipated quality. Such firm-specific capital costs could, for example, take the form of sunk investments in the design of a firm logo or an expensive sign promoting the firm's name. Expenditures on these highly firm-specific assets are then said to represent brand name (or selling) capital investments.

The competitive process also forces the firm-specific capital investments to take the form of assets which provide the greatest direct service value to consumers. The consumers' "effective" price of purchasing a quality-assured good, where the effective price is defined as the purchase price of a product, P^*, less the value of the services yielded by the jointly supplied brand name assets, is thereby minimized. Competition among firms in seeking and making the most highly valued firm-specific brand name capital investments will occur until the expected wealth increase and, therefore, the incentive to enter the industry are eliminated.

If the firm decides to cheat it will experience a capital loss equal to its anticipated future profit stream. Since $P^*(q)$ is derived so that the threat of loss of this future profit stream motivates guaranteed quality, the zero-profit equilibrium "brand-name capital," β, which serves as an explicit hostage to prevent cheating, equals, in terms of figure 1, where $P^*(q_h) = P_2$,

$$\beta = \frac{[P_2 - (AC)_0]x_2}{r}. \tag{7}$$

That is, the market value of the competitive firm's brand name capital is equal to the value of total specific or "sunk" selling costs made by the firm which, in turn, equals the present value of the anticipated premium stream from high quality output. If we continue to assume that there are no capital (and therefore "sunk") costs of production, the

zero-profit equilibrium is shown in figure 1 where average "total" cost (which includes average production costs, AC_{qh}, plus average brand-name capital [i.e., nonsalvageable "selling"] costs, $r[\beta/x]$) just equals price, P_2.

What assures high quality supply is the capital loss due to the loss of future business if low quality is produced. Since the imputed value of the firm's brand name capital is determined by the firm's expected quasi rents on future sales, this capital loss from supplying quality lower than promised is represented by the depreciation of this firm-specific asset. The expenditures on brand name capital assets are therefore similar to collateral that the firm loses if it supplies output of less than anticipated quality and in equilibrium the premium stream provides only a normal rate of return on this collateral asset.

Note that the "effective" price paid by consumers, which equals the quality-assuring price less the value of the consumer services yielded by the brand name capital, may be minimized by the investment in specific selling assets with some positive salvage value. Even though this results in an increased quality-guaranteeing price, assets with positive salvage values may yield differentially large direct consumer service flows. All brand name capital assets must, however, satisfy a necessary condition that the salvage value per unit of output be less than the consumer service value. Firms competing to minimize the effective price will choose specific assets by trading off increased consumer service value with decreased salvage value. This may explain why stores which supply high quality products often have amenities (such as a luxurious carpet cut to fit the particular store) even though only small direct consumer services may be yielded relative to cost.[13]

B. *Nonsalvageable Productive Assets*

The market equilibrium we have developed implies an effective price for high quality output that is higher than what would exist in a zero information cost world. While the costless-information solution is meaningless as an achievable standard of "efficiency," alternative marketing arrangements may be usefully compared to this benchmark. Viable, competitive firms will adopt the arrangements which, considering all transacting and contracting costs, minimize the devia-

[13] If the "sunk" asset yields absolutely no consumer services, then the firm will not use it. Even though profits would be eliminated by purchase of such an asset, consumers would be indifferent between a firm that invested in the asset and a firm that did not. In a world where consumers do not possess full knowledge of cost conditions, however, use of obviously specific assets may be employed even if yielding no direct consumer service flow because they may efficiently inform consumers regarding the sunk capital cost to the firm. This is discussed in greater detail in Sec. IIIC.

tions between the costless-information price and the effective price. One potentially efficient alternative or supplement to the pure price-premium method of guaranteeing quality may be the use of nonsalvageable productive assets rather than brand name (selling) assets.

In order to simplify the analysis of price premiums in guaranteeing quality, we have assumed that all production costs, including fixed costs, were noncapital costs and therefore, by definition, salvageable. More realistically, firms can control both the capital intensity of production and the salvage value of any fixed assets employed in the production process. In particular, if the firm uses a production process that has a nonsalvageable capital element, the normal rate of return (quasi-rent stream) on this element of production capital effectively serves as a quality-assuring premium. In terms of our model, the capital value of the quasi-rent stream when a firm cheats (eq. [3]) is now modified so that the net gain from cheating equals W_3 minus this nonsalvageable capital cost. Alternatively, in the zero-profit equilibrium the total level of collateral must still equal the potential gross gains from cheating, but part of the collateral is now provided by the nonsalvageable production assets rather than the brand name capital assets.

For example, if a machine is somewhat illiquid, buying it rather than renting it short term provides some of this collateral and lowers the quality-guaranteeing price. In fact, because of positive selling costs, capital assets generally have a salvage value less than cost. Thus capital inputs, especially those that have a high turnover cost, will have a value in terms of providing quality assurance in addition to their productive value. Even if the asset is not firm specific, if there is any time delay after the firm cheats and is terminated by consumers in selling the asset to another firm, the firm loses the real rate of interest for that time period on the capital. In addition to physical capital, human capital costs, especially entrepreneurial skills, are also often highly nonsalvageable in the face of consumer termination and therefore also provide significant quality assurance.

The general theoretical point is that the presence of positive quality-information costs favors an increase in the capital intensity of production, including the extent of long-term, illiquid contractual arrangements with suppliers of productive inputs. In particular, the minimum-cost production technique is no longer necessarily that which minimizes solely the average cost of production. "Sunk" production capital now accomplishes two functions—the supply of production services and the supply of quality-assuring services. Therefore, increases in average production costs accompanied by larger increases in sunk production assets may minimize the effective consumer product price. Profit maximization requires firms to trade off "inefficient"

production technologies and the quality-assurance cost savings implied by the presence of firm-specific (sunk) capital assets in the productive process and hence the reduced necessity for the firm to make sunk selling cost (brand name capital) investments. Although the more capital intensive production technology may increase the perfectly competitive price of high quality output, P_0, it reduces the price premium, $P_2 - P_1$, necessary to assure the supply of that high quality. In fact, even a very slight modification of the minimum production cost technology, such as an alteration in some contractual terms, may imply the existence of large enough nonsalvageable assets so that the need for a quality-guaranteeing price premium is eliminated entirely.[14]

C. *Consumer Cost Uncertainty: A Role for Advertising*

The discussion to this point has assumed complete consumer knowledge of firms' costs of producing alternative quality outputs and knowledge of the extent to which any capital production costs or brand name capital selling costs are salvageable. This knowledge is necessary and sufficient to accurately calculate both the quality-guaranteeing premium and price. However, consumers are generally uncertain about cost conditions and therefore do not know the minimum quality-guaranteeing price with perfect accuracy. In fact, consumers cannot even make perfect anticipated quality rankings across firms on the basis of price. That one firm has a higher price than another may indicate a larger price premium or, alternatively, more inefficient production. In this section, we examine how the more realistic assumption of consumer cost uncertainty influences market responses to prepurchase quality uncertainty.

We have shown that increases in the price premium over average recoverable cost generally increase the relative returns from production of promised (high) quality rather than deceptive minimum (low) quality. The existence of a high price premium also makes expenditures on brand name capital investments economically feasible. The magnitude of brand name capital investments in turn indicates

[14] For example, franchisers can assure quality by requiring franchisee investment in specific production capital. A general arrangement by which this is accomplished is by not permitting the franchisee to own the land upon which its investments (e.g., capital fixtures) are made. Rather, the franchiser owns or leases the land and leases or subleases it to the franchisee, thereby creating for the franchisee a large nonsalvageable asset if he is terminated by the franchiser. This highly franchiser-specific asset can therefore serve as a form of collateral and potentially eliminate any need for a price premium. See Klein (1980) for a more complete discussion of this franchising solution, including the potential reverse cheating problem that is created by such contractual arrangements.

the magnitude of the price premium. When a consumer is uncertain about the cost of producing a particular high quality level of output and therefore the required quality-assuring premium, information about the actual level of the price premium will provide information about the probability of receiving high quality. If consumers are risk averse, this uncertainty about receiving anticipated high or deceptively low quality output will increase the premium that will be paid. The premium will include both a (presumably unbiased) estimate of the quality-assuring premium and an extra payment to reduce the risk of being deceived.

Thus, when consumers do not know the minimum quality-guaranteeing price, the larger is a firm's brand name capital investment relative to sales, the more likely its price premium is sufficient to motivate high quality production. Competitive investment in brand name capital is now no longer constrained to assets which yield direct consumer service flows with a present discounted value greater than the salvage value of the assets. Implicit information about the sufficiency of price as a guarantee can be supplied by "conspicuous" specific asset expenditures. Luxurious storefronts and ornate displays or signs may be supplied by a firm even if yielding no direct consumer service flows. Such firm-specific assets inform consumers of the magnitude of sunk capital costs and thereby supply information about the quasi-rent price-premium stream being earned by the firm and hence the opportunity cost to the firm if it cheats. Both the informational services and the direct utility producing services of assets are now relevant considerations for a firm in deciding upon the most valuable form the brand name capital investment should take.

The value of information about the magnitude of a firm's specific or "sunk" capital cost, and therefore the magnitude of the price premium, is one return from advertising. Indeed, the role of premiums as quality guarantors provides foundation for Nelson's (1974) argument that advertising, by definition, supplies valuable information to consumers—namely, information that the firm is advertising. A sufficient investment in advertising implies that a firm will not engage in short-run quality deception since the advertising indicates a nonsalvageable cost gap between price and production costs, that is, the existence of a price premium. This argument essentially reverses Nelson's logic. It is not that it pays a firm with a "best buy" to advertise more, but rather that advertising implies the supply of "best buys," or more correctly, the supply of promised high quality products. Advertising does not directly "signal" the presence of a "best buy," but "signals" the presence of firm-specific selling costs and therefore the magnitude of the price premium. We would therefore expect, ceteris paribus, a positive correlation not between advertising intensity and

"best buys," as Nelson claims, but between advertising intensity and the extent of quality that is costly to determine prepurchase.[15]

Conspicuous sunk costs such as advertising are, like all sunk costs, irrelevant in determining future firm behavior regarding output quality. However, consumers know that such sunk costs can be profitable only if the future quasi rents are large. In particular, if the consumer estimate of the initial sunk expenditure made by the firm is greater than the consumer estimate of the firm's possible short-run cheating gain, then a price premium on future sales sufficient to prevent cheating is estimated to exist. Our analysis therefore implies that independent of excludability or collection costs, advertising that guarantees quality will be sold at a zero price and "tied in" with the marked-up product being advertised.[16]

Our theory also suggests why endorsements by celebrities and other seemingly "noninformative" advertising such as elaborate (obviously costly to produce) commercials, sponsorships of telethons, athletic events, and charities are valuable to consumers. In addition to drawing attention to the product, such advertising indicates the presence of a large sunk "selling" cost and the existence of a price premium. And because the crucial variable is the consumers' estimate of the stock of advertising capital (and not the flow), it also explains why firms advertise that they have advertised in the past (e.g., "as seen on 'The Tonight Show'"). Rather than serving a direct certifying function (e.g., as recommended by *Good Housekeeping* magazine), information about past advertising informs consumers about the magnitude of the total brand name capital investment.[17]

Firms may also provide valuable information by publicizing the

[15] Nelson's argument is based on an assumption similar to the Spence (1973)-type screening assumption regarding the lower cost to more productive individuals of obtaining education. Nelson's argument, however, is circular since consumers react to advertising only because the best buys advertise more and the best buys advertise more only because consumers buy advertised products. Schmalensee (1978) has shown that the Nelson scenario may imply "fly-by-night" producers who advertise the most and also deceptively produce minimum quality. Like Spence's signaling model, the government could, in principle, tax this investment and thereby save real resources without reducing the effectiveness of this information if consumers were aware of the tax rate. However, advertising serves many purposes. In particular, advertising also can supply valuable consumer information about the particular characteristics and availability of products. For optimality the government would therefore have to determine the appropriate tax rate for each advertising message and consumers would have to be aware of each of these particular tax rates.

[16] Mishan (1970) has argued for legislation which would require advertising to be sold separately at a price which covers advertising costs. This would completely destroy the informational value of advertising we are emphasizing here.

[17] Note, however, that just as firms may deceive consumers about quality to be supplied, they may also attempt to deceive them about the magnitude of the advertising investments made, e.g., purchasing a local spot on "The Tonight Show" and advertising the advertising as if an expenditure on a national spot was made.

large fees paid to celebrities for commercials. Information about large endorsement fees would be closely guarded if the purpose were to simulate an "unsolicited endorsement" of the product's particular quality characteristics rather than to indicate the existence of a price premium. Viewed in this context, it is obviously unnecessary for the celebrity to actually use the particular brand advertised. This is contrary to a recent FTC ruling (see Federal Trade Commission 1980).

This analysis of advertising implies that consumers necessarily receive something when they pay a higher price for an advertised brand. An expensive name brand aspirin, for example, is likely to be better than unadvertised aspirin because it is expensive. The advertising of the name brand product indicates the presence of a current and future price premium. This premium on future sales is the firm's brand name capital which will be lost if the firm supplies lower than anticipated quality. Therefore, firms selling more highly advertised, higher priced products will necessarily take more precautions in production.[18]

We have emphasized the informational value of advertising as a sunk cost. Other marketing activities can serve a similar informational role in indicating the presence of a price premium. For example, free samples, in addition to letting consumers try the product, provide information regarding future premiums and therefore anticipated quality. Such free or low-price samples thus provide information not solely to those consumers that receive the samples but also to anyone aware of the existence and magnitude of the free or low-price sample program. More generally, the supply by a firm of quality greater than anticipated and paid for by consumers is a similar type of brand name capital investment by the firm. By forgoing revenue, the firm provides information to consumers that it has made a nonsalvageable investment of a particular magnitude and that a particular future premium stream is anticipated to cover the initial sunk alternative cost.[19]

[18] The greater is the cost to consumers of obtaining deceptively low quality, the greater will be the demand for quality assurance. The very low market share of "generic" children's aspirin (1 percent) vis-à-vis generic's share of the regular aspirin market (7 percent) is consistent with this implication (see IMS America, Ltd. 1978). Many individuals who claim "all aspirin is alike" apparently pay the extra price for their children where the costs of lower quality are greater and therefore quality assurance is considered more important.

[19] Our analysis of advertising also illuminates the monopolistic competition debate. Chamberlin's (1965) distinction between production costs, defined as what is included in the "package" that passes from seller to buyer, and selling costs (e.g., advertising), which are not part of the package transferred, suggests that selling costs usefully may be considered as a privately supplied collective factor. For example, a firm which holds selling costs, such as expenditures on a store sign, constant as his sales increase does not appear to be decreasing the average "quality" of his product. Demsetz (1959, 1968) made the contrary assumption that average quality does fall as sales increase, holding

Finally, even when consumers systematically underestimate the quality-assuring price because of downward-biased estimates of production or marketing costs or upward-biased estimates of anticipated demand growth, firms in a monopolistically competitive environment may not cheat. Such price-setting firms may possess specific nonsalvageable assets (such as trademarks) upon which they are earning a sufficient quasi-rent premium to induce high quality supply. However, the existence of independent competitive retailers that do not have any ownership stake in this firm-specific asset and yet can significantly influence the quality of the final product supplied to consumers creates a severe quality-cheating problem for the manufacturer. In this context, rational but imperfectly informed consumers will not demand a sufficient premium to prevent retailer cheating. Manufacturers may protect their trademarks by imposing constraints on the retailer competitive process including entry restrictions, exclusive territorial grants, minimum resale price maintenance, and advertising restrictions that will assure quality by creating a sufficiently valuable premium stream for the retailers. If this manufacturer-created premium stream is greater than the potential short-run retailer return from deceptive low quality supply, the magnitude of which is determined in part by the manufacturer by its level of direct policing expenditures, the retailer will not cheat and the consumer will receive anticipated high quality supply.[20]

IV. Conclusion

We have shown that even the existence of perfect communication among buyers so that all future sales are lost to a cheating firm is not sufficient to assure noncheating behavior. We have analyzed the

---

selling costs constant, by merely ignoring Chamberlin's distinction and its possible theoretical significance and identifying quality costs with selling costs (aggregating both into the concept "demand increasing costs"). However, since in a monopolistically competitive environment the price premium that will assure quality depends upon the demand expected in the future, the quality incentive implied by an advertising investment also depends upon consumers' expectations about future demand. In particular, the relevant variable indicating an incentive to produce high quality is the level of advertising capital compared to anticipated future sales. Hence advertising is not a pure public good in a firm's production function as Chamberlin implicitly assumed, and the arbitrary contrary assumption made by Demsetz is possibly justifiable.

[20] See Klein et al. (1980) for a complete analysis of this case applied to the FTC Coors litigation. Coors appears to have employed exclusive territories on the wholesale level and resale price maintenance on the retail level to create a sufficient premium to encourage the necessary refrigeration of their nonpasteurized beer. Implications of this analysis in terms of providing a possible rationale for similar constraints on the competitive process enforced by trade associations and government regulatory agencies are also examined.

generally unrecognized importance of increased market prices and nonsalvageable capital as possible methods of making quality promises credible. We obviously do not want to claim that consumers "know" this theory in the sense that they can verbalize it but only that they behave in such a way as if they recognize the forces at work. They may, for example, know from past experience that when a particular type of investment is present such as advertising they are much less likely to be deceived. Therefore, survivorship of crude decision rules over time may produce consumer behavior very similar to what would be predicted by this model without the existence of explicit "knowledge" of the forces we have examined.

Our analysis implies that consumers can successfully use price as an indicator of quality. We are not referring to the phenomenon of an ignorant consumer free riding on the information contained in the market price paid by other more informed buyers but rather to the fact that consumer knowledge of a gap between firm price and salvageable costs, that is, the knowledge of the existence of a price premium, supplies quality assurance. The former argument, that a naive buyer in a market dominated by knowledgeable buyers can use price as a quality signal because the relative market price of different products reflects differences in production costs and therefore differences in quality, crucially depends upon a "majority" of the buyers in the market being knowledgeable.

As Scitovsky (1945, p. 101) correctly notes, ". . . the situation becomes paradoxical when price is the index by which the average buyer judges quality. In a market where this happens price ceases to be governed by competition and becomes instead an instrument wherewith the seller can influence his customer's opinions of the quality of his wares." However, even when the "average" buyer uses price as an index of quality, we need not fear, as Scitovsky does, the havoc this supposedly wreaks on the economic theory of choice. All consumers in a market may consistently use price, given their estimates of salvageable production costs, as an indicator of the firm's price-premium stream and therefore as an indicator of the anticipated quality of the output to be supplied by the firm. Scitovsky did not consider that price not only influences buyers' expectations but also influences producers' incentives.

We do not wish to suggest that use of implicit (price premium–specific investment) contracts is always the cheapest way to assure quality supply. When quality characteristics can be specified cheaply and measured by a third party, and hence contract enforcement costs are anticipated to be low, explicit contractual solutions with governmentally enforced penalties (including warranties) may be a less

costly solution. When explicit contract costs are high and the extent of short-run profit from deceptively low quality supply and hence the quality-assuring price premium is also high, governmental specification and enforcement of minimum quality standards may be an alternative method of reducing the costs of assuring the supply of high quality products.[21] And, finally, vertical integration, which in this consumer-product context may consist of home production or consumer cooperatives, may be a possible alternative arrangement (see Klein, Crawford, and Alchian 1978).

The three major methods in which to organize transactions can be usefully considered within this framework as (*a*) explicit contractual or regulatory specification with third-party enforcement, (*b*) direct (two-party) enforcement of implicit contracts, and (*c*) one-party organization or vertical integration. This paper has analyzed the brand name repeat-purchase mechanism represented by the second alternative. More generally, however, all market transactions, including those "within" the firm such as employer-employee agreements, consist of a combination of the two basic forms of contractual arrangements. Some elements of performance will be specified and enforced by third-party sanctions and other elements enforced without invoking the power of some outside party to the transaction but merely by the threat of termination of the transactional relationship.

Our analysis implies that, given a particular level of explicit contract costs, we are more likely to observe an increased reliance on the brand name contract-enforcement mechanism the lower the rate of interest and the lower the level of prepurchase quality-determination costs. The lower the interest rate the greater the capital cost to a firm from the loss of future sales and therefore the lower the equilibrium price premium. Hence we can expect the termination of future exchange method of enforcing contracts to be more effective. More generally, since the interest rate in our model refers to the period of product repurchase, the quality assurance will be less costly for less durable goods that have greater repurchase frequency. Franchising chains, for example, take advantage of this effect by making it possible for

[21] Such governmental regulations, however, do not avoid the contractual problems of ex ante explicitly defining in an enforceable manner all major elements of performance. Nor do they necessarily avoid the implicit contractual conditions of a price-premium stream (created by entry restrictions, an initial forfeitable bond, and/or minimum price restraints) to effectively enforce the governmental regulations (see Klein et al. 1980). In addition, by making it illegal to supply less than the regulated quality, individuals that would voluntarily demand lower quality than the regulated standard incur a loss of consumer surplus. Distribution effects are created, since while the regulation may decrease the cost of supplying high quality output it increases the cost of supplying lower quality output

consumers to pool information from sales of seemingly disparate sellers, thereby decreasing the period of repurchase and the quality-assuring price.

Similarly, purchase from a diversified firm increases the frequency of repeat purchase and lowers the necessary price premium. As long as consumers react to receiving unexpectedly low quality from a diversified firm by reducing purchases of the firm's entire product line, all the firm's nonsalvageable capital serves to assure the quality of each product it produces. This economy of scale in communicating quality-assurance information to consumers may be one motivation for conglomerate mergers. If a firm sells a set of products, each of which is produced by capital with salvage value less than costs, the quality-guaranteeing price premium on each product will be lower than if production were done by separate firms.

Finally, we can expect greater reliance on the non-third-party method of contract enforcement the lower the direct costs to the consumer of determining quality of the product prepurchase. The higher the costs of producing the minimum quality output that cannot be distinguished prepurchase from a given promised high quality output and the faster these minimum quality production costs rise with increased output, the lower the potential short-run cheating gain and therefore the lower the price premium. When the low quality cost function is such that a cheating firm can expand output a substantial amount with little increase in cost, use of the brand name enforcement mechanism is unlikely.

When the low quality cost function becomes so flat that the premium solution does not exist, the implicit contract-enforcement mechanism we have analyzed will not be used. When this condition is combined with an extremely high cost of quality assurance via explicit contractual guarantees, governmental supply may be the cheapest alternative. An obvious example is the good "money," where the marginal cost of production is essentially zero, the short-run cheating potential extremely large, and where the cost of a commodity money or the necessary bullion reserves to assure performance via convertibility is also extremely high. Governmental supply is the generally adopted but far from costless solution (see Klein 1974). Other products where the "hold-up" potential is very large and where explicit contract costs are high (such as police or fire protection services) are also generally supplied by non-profit-maximizing government agencies rather than by unregulated profit-maximizing firms earning large quasi rents on unsalvageable (firm-specific) capital assets. In general, minimization of the cost of assuring performance will imply an optimal combination of governmental regulation and/or supply, explici

contractual enforcement, vertical integration, and the implicit (brand name) contractual enforcement mechanism we have analyzed.

Appendix

Simulation of the Quality-assuring Price

If we assume that output of high and low (minimum) quality is produced by constant-elasticity cost functions of the form:

$$C_h = F_h + \beta_h x_h^{\alpha}, \tag{A1}$$

$$C_l = F_l + \beta_l x_l^{\alpha}, \tag{A2}$$

the quality-assuring price premium, $\tilde{P}$, will be given by:[22]

$$\tilde{P} = \left\{ \frac{\left[1 - \frac{(F_l/F_h)r}{1+r}\right]}{1 - \frac{r}{1+r}\left(\frac{\beta_l}{\beta_h}\right)^{1/(\alpha-1)}} \right\}^{(\alpha-1)/\alpha}. \tag{A3}$$

This expression indicates that as the ratio of low to high quality fixed costs, (F_l/F_h), decreases the quality-guaranteeing price-premium increases (because the short-run profit from cheating increases). But, as long as F_l can be assumed to be less than or equal to F_h, fixed costs cannot affect the existence of the quality-guaranteeing price premium. Similarly, as the interest rate, r, increases, the quality-guaranteeing price premium increases but will always exist. It is the marginal cost elasticity, $[1/(\alpha - 1)]$, and the ratio of the marginal cost slopes, (β_l/β_h), that determine the existence of a quality-assuring price premium. An increase in the elasticity of marginal cost or a decrease in the ratio of the low to high quality marginal cost slopes, by increasing the possible expansion of the low quality output at the high quality-guaranteeing price, increases the quality-guaranteeing price premium and the likelihood that it may not exist. Simulation results as a function of these parameters are presented in table A1 below. The ratio of the quality-assuring price to the minimum average cost of high quality production, P_2/P_1 in terms of figure 1, along with the ratio of low quality output at the quality-assuring price relative to the minimum high quality average cost output, x_4/x_1 in terms of figure 1, is presented. When the quality-assuring price does not exist, the ratio of low quality output at a P_2/P_1 ratio of 2 to the output at the minimum average cost of low quality output, x_0, is presented in brackets to indicate the shape of the low quality cost function. The results indicate that these cases of nonexistence generally occur where the low quality cost curve is so flat relative to the high quality cost curve that cheating output can be expanded dramatically relative to the noncheating output. For example, when the marginal cost elasticity is assumed to be 10.0 and the ratio of marginal cost slopes is assumed to be 0.25, it implies that low quality output can be profitably expanded by more than a billion times beyond its minimum average cost rate when the market price is double the perfectly competitive high quality price.

[22] The derivation is available to readers upon request.

TABLE A1

SIMULATION OF QUALITY-ASSURING PRICES

Interest Rate r	Ratio of Fixed Costs (F_l/F_h)	Marginal Cost Elasticity $[1/(\alpha - 1)]$	Ratio of Marginal Cost Slopes (β_l/β_h)	Price-Premium Ratio (P_2/P_1)	Output Ratio (x_4/x_1)
.03	.5	5	.25	1.031	2.03
.03	.5	.5	.50	1.018	1.43
.03	.5	.5	.75	1.013	1.16
.03	.5	2.0	.25	1.227	24.07
.03	.5	2.0	.50	1.037	4.30
.03	.5	2.0	.75	1.013	1.82
.03	.5	4.0	.25	Does not exist	$[4.10 \times 10^3]$
.03	.5	4.0	.50	1.130	26.12
.03	.5	4.0	.75	1.017	3.37
.03	.5	10.0	.25	Does not exist	$[1.07 \times 10^9]$
.03	.5	10.0	.50	Does not exist	$[1.05 \times 10^6]$
.03	.5	10.0	.75	1.067	33.97
.03	1.0	.5	.25	1.021	2.02
.03	1.0	.5	.50	1.008	1.42
.03	1.0	.5	.75	1.003	1.16
.03	1.0	2.0	.25	1.221	23.84
.03	1.0	2.0	.50	1.032	4.26
.03	1.0	2.0	.75	1.008	1.81
.03	1.0	4.0	.25	Does not exist	$[4.10 \times 10^3]$
.03	1.0	4.0	.50	1.127	25.81
.03	1.0	4.0	.75	1.013	3.34
.03	1.0	10.0	.25	Does not exist	$[1.07 \times 10^9]$
.03	1.0	10.0	.50	Does not exist	$[1.05 \times 10^6]$
.03	1.0	10.0	.75	1.066	1.82×10^4
.09	.5	.5	.25	1.097	2.09
.09	.5	.5	.50	1.056	1.45
.09	.5	.5	.75	1.039	1.18

.09	.5	2 0	.25	Does not exist	[64.0]
.09	.5	2 0	.50	1.127	5.08
.09	.5	2.0	.75	1.040	1.92
09	5	4 0	25	Does not exist	[4.10×10^3]
.09	.5	4.0	.50	Does not exist	[2.56×10^2]
.09	.5	4.0	.75	1.053	3.89
.09	.5	10.0	25	Does not exist	[1.07×10^9]
.09	.5	10 0	.50	Does not exist	[1.05×10^6]
.09	5	10.0	.75	Does not exist	[1.82×10^4]
09	1.0	.5	.25	1.065	2.06
.09	1.0	.5	50	1.026	1.43
.09	1.0	5	.75	1.009	1.16
.09	1.0	2.0	.25	Does not exist	[64.0]
09	1 0	2.0	.50	1.111	4.93
09	1.0	2.0	.75	1.024	1.87
09	1.0	4.0	25	Does not exist	[4.10×10^3]
.09	1.0	4.0	.50	Does not exist	[2.56×10^2]
.09	1.0	4 0	.75	1.044	3.76
.09	1 0	10.0	.25	Does not exist	[1.07×10^9]
09	1.0	10.0	.50	Does not exist	[1.05×10^6]
.09	1 0	10.0	.75	Does not exist	[1.82×10^4]

References

Akerlof, George A. "The Market for 'Lemons': Quality Uncertainty and the Market Mechanism." *Q.J.E.* 84 (August 1970): 488–500.

Becker, Gary S., and Stigler, George J. "Law Enforcement, Malfeasance, and Compensation of Enforcers." *J. Legal Studies* 3 (January 1974): 1–18.

Chamberlin, Edward H. *The Theory of Monopolistic Competition: A Re-Orientation of the Theory of Value.* 8th ed. Cambridge, Mass.: Harvard Univ. Press, 1965.

Darby, Michael R., and Karni, Edi. "Free Competition and the Optimal Amount of Fraud." *J. Law and Econ.* 16 (April 1973): 67–88.

Demsetz, Harold. "The Nature of Equilibrium in Monopolistic Competition." *J.P.E.* 67, no. 1 (February 1959): 21–30.

———. "Do Competition and Monopolistic Competition Differ?" *J.P.E.* 76, no. 1 (January/February 1968): 146–48.

Federal Trade Commission, Office of the Federal Registrar. *Guides concerning Use of Endorsements and Testimonials in Advertising.* 16 CFR, pt. 255. Washington: Government Printing Office, 1980.

Goldberg, Victor P. "Toward an Expanded Theory of Contract." *J. Econ. Issues* 10 (March 1976): 45–61.

Hayek, Friedrich A. "The Meaning of Competition." In *Individualism and Economic Order.* Chicago. Univ. Chicago Press, 1948.

Hobbes, Thomas. *Leviathan.* Oxford: Blackwell, 1955, first published 1651.

IMS America, Ltd. *U.S. Pharmaceutical Market, Drug Stores and Hospitals, Audit of Purchases.* Bergen, N.J.: IMS America, 1978.

Klein, Benjamin. "The Competitive Supply of Money." *J. Money, Credit and Banking* 6 (November 1974): 423–53.

———. "Borderlines of Law and Economic Theory: Transaction Cost Determinants of 'Unfair' Contractual Arrangements." *A.E.R. Papers and Proc.* 70 (May 1980): 356–62.

Klein, Benjamin; Crawford, Robert G.; and Alchian, Armen A. "Vertical Integration, Appropriable Rents, and the Competitive Contracting Process." *J. Law and Econ.* 21 (October 1978): 297–326.

Klein, Benjamin; McLaughlin, Andrew; and Murphy, Kevin M. "Resale Price Maintenance, Exclusive Territories, and Franchise Termination: The Coors Case." Working Paper, UCLA, 1980.

Leffler, Keith B. "The Role of Price in Guaranteeing Quality." Working Paper no. CPB77-5, Univ. Rochester, June 1977.

Luce, R. Duncan, and Raiffa, Howard. *Games and Decisions: Introduction and Critical Survey.* New York: Wiley, 1957.

Macaulay, Stewart. "Non-contractual Relations in Business: A Preliminary Study." *American Soc. Rev.* 28 (February 1963): 55–67.

Macneil, Ian. "The Many Futures of Contracts." *Southern California Law Rev.* 47 (May 1974): 691–816.

Marshall, Alfred. *Principles of Economics: An Introductory Volume.* 8th ed. New York: Macmillan, 1949.

Mishan, Edward J. *21 Popular Economic Fallacies.* New York: Praeger, 1970.

Nelson, Phillip. "Information and Consumer Behavior." *J.P.E.* 78, no. 2 (March/April 1970): 311–29.

———. "Advertising as Information." *J.P.E.* 82, no. 4 (July/August 1974): 729–54.

Schmalensee, Richard. "A Model of Advertising and Product Quality." *J.P.E.* 86, no. 3 (June 1978): 485–503.

Scitovsky, Tibor. "Some Consequences of the Habit of Judging Quality by Price." *Rev. Econ. Studies* 12, no. 2 (1945): 100–105.

Smith, Adam. *An Inquiry into the Nature and Causes of the Wealth of Nations*. New York: Modern Library, 1937, first published 1776.

Spence, A. Michael. "Job Market Signaling." *Q.J.E.* 87 (August 1973): 355–74.

Telser, Lester G. "A Theory of Self-enforcing Agreements." *J. Bus.* 22, no. 1 (January 1980): 27–44.

Williamson, Oliver E. "Transaction-Cost Economics: The Governance of Contractual Relations." *J. Law and Econ.* 22 (October 1979): 233–61.

The Political Economy of Benefits and Costs: A Neoclassical Approach to Distributive Politics

Barry R. Weingast, Kenneth A. Shepsle, and Christopher Johnsen

Washington University

This essay offers a rational political explanation for the notorious inefficiency of pork barrel projects with an optimization model of legislative behavior and legislative institutions. The model emphasizes the (economically arbitrary, from a welfare point of view) importance of the geographic incidence of benefits and costs owing to the geographic basis for political representation. We explore the implications of a legislator's objective function and derive conditions under which a representative legislature will select an omnibus of projects each of which exceeds the efficient scale.

The inefficiency of public decision making has long been a concern for students of public policy. Models of bureaucratic behavior, legislative institutions, interest-group influence, vote-maximizing politicians, and fiscal illusion all point toward sources of bias in these nonmarket contexts.[1] Yet none has established why a cooperative legislature would stand for policies which are Pareto dominated.

Mr Weingast is assistant professor of Economics and research associate at the Center for the Study of American Business. Mr. Shepsle is professor of Political Science and research associate of the Center for the Study of American Business Mr. Johnsen is a doctoral candidate in the Department of Political Science. The authors thank Roger Noll of the California Institute of Technology for a timely conversation. They are also grateful for the constructive comments of Morris Fiorina, George Stigler, and an anonymous referee

[1] Consult Downs 1957, 1967; Niskanen 1971; Riker and Brams 1973; Breton 1974; Ferejohn and Fiorina 1975; Wagner 1976; Aranson and Ordeshook 1978; and Fiorina and Noll 1978.

[*Journal of Political Economy*, 1981, vol 89, no 4]
© 1981 by The University of Chicago 0022-3808/81/8904-0008$01 50

Specifically, if the bias works to the benefit of some identifiable group, then why do political actors not insist on efficiency in combination with a compensation scheme? This would leave the identified group at least as well off, all others at least as well off, and then leave the efficiency gain at the political discretion of representatives, thereby enhancing their role. From the anomaly of sustained political inefficiency, it appears that there is a divergence between normative economic principles, on the one hand, and the preferences of political actors on the other. Prominence is given in the following analysis to the political mechanisms that create and maintain this divergence.

In this paper, we develop a model of the public choice mechanisms comprising a representative legislature in order to show the political sources that systematically bias public decisions toward larger than efficient projects in the area of distributive policymaking. By distributive policies we mean those projects, programs, and grants that concentrate the benefits in geographically specific constituencies, while spreading their costs across all constituencies through generalized taxation.[2] This collection of public decisions includes the traditional pork barrel of public works and rivers and harbors projects as well as the more recent examples of highway construction, categorical grants-in-aid, urban renewal, mass transit, and sewage treatment plants. The model identifies the political sources of efficiency bias—and there are several—by unpacking the democratic institutions into their components, thereby focusing on the influence of each. This takes the form of a progression of models beginning with an efficiency benchmark. Then, one by one, political features are added until the final form models a representative legislature divided into n districts. The approach shows how political institutions transform the economic basis of costs and benefits into political costs and benefits. The latter, and not their economic counterparts, define rational decisions for political actors.

The model reveals three important sources of bias. The first is a consequence of the political definition of benefits and costs and its divergence, in important respects, from the economic definition. The second source stems from the districting mechanism which divides the economy into n disjoint political units called districts. The method of project financing through generalized taxation constitutes the third source of bias. Moreover, we show that the mechanism of popular election of legislative representatives complements these sources of bias so that these three sources, in conjunction with the reelection mechanism, explain the inefficiency of political choice.

[2] See Lowi (1964) for a discussion of different kinds of policies based on the characteristics of beneficiaries as well as the mechanisms of financing.

The General Approach

Our model is one of policymaking in the realm of distributive policy, so we begin by clarifying that concept. A distributive policy is a political decision that concentrates benefits in a specific geographic constituency and finances expenditures through generalized taxation. These policies—sewage treatment plants, land reclamation, the rivers and harbors omnibus, urban renewal projects—authorize collections of projects, each targeted to a geographic location and each generating benefits in that geographic location unrelated to projects in other locations. This latter characteristic is crucial since the omnibus-like quality of distributive programs allows decisions to be made on a project-by-project, locality-by-locality basis. Each project of the omnibus may be fashioned independent of others in the omnibus.

While it is clear that all policies have a geographic incidence of benefits and costs, what distinguishes a distributive policy is that benefits are geographically targeted. In contrast, a nondistributive program, say an entitlement program, though having geographic incidences, is fashioned with nongeographic constituencies in mind, for example, socioeconomic groups. Subsidies to beekeepers, for example, generate a distribution of benefits that depends on the geographic distribution of beekeepers. No geographic area has a claim on program benefits except as it contains residents in the entitled category. Thus, by our definition, programs targeted to the malnourished (food stamps), the unhealthy (Medicare), the poor (welfare), the retired (social security), the injured worker (workmen's compensation), or the automobile driver (automotive product safety) are not distributive policies because any citizen may obtain program benefits if he falls in the specific category. Thus, an entitlement program confers benefits on all individuals in the designated category, benefits which may not be varied without similarly varying them for others in the category. Although the motivation to create nondistributive programs may have a geographic basis (as when a politician is moved to support a policy because many of his constituents fall into the beneficiary group), the fact remains that the beneficiary group is not geographically defined or determined. In contrast, geography is the hallmark of distributive politics: Programs and projects are geographically targeted, geographically fashioned, and may be independently varied. Importantly, geography is also the basis for political organization and representation.

A distributive policy for the jth district, $P_j(x)$, is a project located in that district, where x is a decision parameter. Although x may be treated as a vector of project characteristics, we assume that x simply describes the scale or size of the project. Associated with the project

$P_j(x)$ are both benefits and costs. Let $b(x)$ represent the present value of the economic benefits which flow from the project to the particular political constituency. This includes consumption benefits, say cleaner water from a sewage treatment plant, and potential pecuniary gains to producers, for example, increased profits to project input owners from price rises in factor markets.[3]

ASSUMPTION 1: $b'(x) > 0$, $b''(x) < 0$.

Let $c(x)$ represent the total resource cost involved in producing the project. It decomposes into three components, $c(x) = c_1(x) + c_2(x) + c_3(x)$. The first component, $c_1(x)$, is the real resource expenditures for project inputs spent in the constituency in which the project is located; $c_2(x)$ is the real resource expenditures for project inputs spent outside the district; $c_3(x)$ is nonexpenditure real resource costs imposed on the district (e.g., nonpecuniary externalities, such as the destruction of the natural environment, and pecuniary externalities in the form of price rises to consumers in factor markets).[4]

ASSUMPTION 2: $c_i'(x) > 0$, $c_i''(x) \geq 0$, $i = 1, 2, 3$.

The expenditures are financed through taxes so that the tax bill for $P_j(x)$ is

$$T(x) = c_1(x) + c_2(x). \tag{1}$$

We assume a tax system that covers all expenditures, assigning nonnegative tax share t_i to the ith district, where $\Sigma_{i=1}^{n} t_i = 1$ and n is the number of districts. The tax bill for the ith district for the project $P_j(x)$, therefore, is $t_i[c_1(x) + c_2(x)]$.

As the above development suggests, there are several mechanisms at work in the realm of distributive policy which our model captures. First, economic benefits are geographically concentrated in a politically relevant way. Second, production costs may be unpacked, again in a politically relevant way. Some costs are extracted from the economy and returned as geographically earmarked expenditures—$c_1(x)$ and $c_2(x)$; other costs are nonexpenditure in nature, imposed on the local economy in which the project is located—$c_3(x)$. Third, the tax bill, $T(x)$, is paid for by each political subdivision according to the tax shares t_i, $i = 1, \ldots, n$. The relevant mechanisms which we examine

[3] Some students of cost-benefit analysis, e.g., McKean (1958, chap. 8), argue against including pecuniary external effects in the calculations, claiming instead that these constitute distributional effects that are not germane for efficiency determinations. The treatment of pecuniary external effects remains controversial in cost-benefit analysis. Our results reported in this paper do not depend on how this controversy is resolved.

[4] There is a fourth component of costs, namely, nonexpenditure real resource costs that spill over into other political constituencies. Since $P_j(x)$ is a concentrated project, we ignore this for the present. We return to this point in the concluding section. For analysis of other kinds of policies not possessing the particular properties of distributive projects, see Shepsle and Weingast (1980).

below are (1) the political cost-accounting mechanism, (2) the districting mechanism, and (3) the taxation mechanism. After examining these mechanisms and their effects on the characteristics of distributive policy, we explore the complementarity of legislators' reelection objective. First, however, we develop the familiar efficiency criterion.

Model E: Maximizing Economic Efficiency

The benchmark for the entire set of political institutions developed below is the efficiency criterion. This requires the maximization of economic net benefits. This is given simply by

$$\max_{x} E(x) = b(x) - c(x). \tag{2}$$

The familiar first- and second-order conditions are

$$b' - c' = 0,$$

and (3)

$$b'' - c'' < 0.$$

The second-order condition follows directly from assumption 1 and assumption 2 so that the solution to (2), x^E in figure 1, is a unique global maximum.

The Politicization of Expenditures

The first transformation of the standard approach is the politicization of economic costs. This transformation rests on a crucial political property. Project costs, paid from general revenues, become geographically earmarked expenditures. The political process distributes those expenditures in the form of c_1-type and c_2-type costs. Thus, for $P_j(x)$ in district j, production inputs are purchased from firms and individuals in the district (c_1) and from their counterparts outside the district (c_2). More important is the political evaluation of the distribution of geographically earmarked expenditures. While these expenditures are not lump-sum transfers of wealth to factor owners, they do entail pecuniary gains since they represent increases in demand for project input factors, thereby driving up their prices.[5] Not only does

[5] This statement presumes that factors of production are geographically fixed in the short run. This does not preclude the bidding away of pecuniary gains as the long-run supply of factors adjusts to this change in demand. For some projects, which are one shot and nonrepetitive in nature, the short-run analysis holds since the long-run adjustment process is attenuated. For other programs, in which a permanent increase in demand has occurred (e.g., the continuing flow of military procurement projects to a district), long-run market forces adjust with the concomitant bidding away of pecuniary

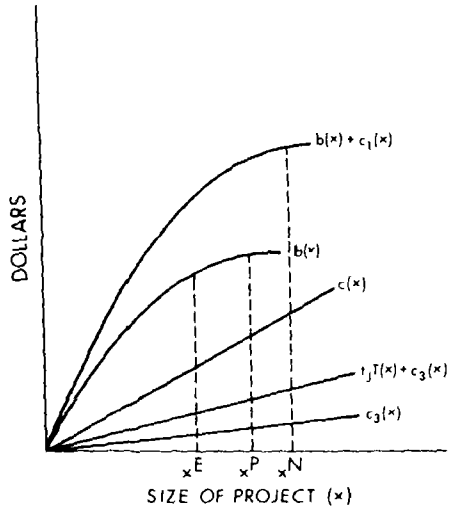


FIG. 1

the public project entail new business for input owners; it allows them to receive a higher price for the sale of inframarginal units as well.[6] As this discussion suggests, geographic expenditures are important because they distribute pecuniary effects. Consequently, the distributional effects of local expenditures combine with consumption benefits in the district's valuation of a project. Put simply, a dollar's valuation of a project may come in either of two forms: a pecuniary gain to a factor owner or a benefit to a project consumer. Partly as a consequence of these distributional effects, and partly for additional reasons enumerated below, the political evaluation of pecuniary effects diverges from their economic treatment.

We may distinguish several classes of agents who are differentially advantaged or harmed (in addition to their tax obligations) by the provision of $P_j(x)$: (1) *in-district consumers*, who receive benefits through consumption of the public project but are unaffected by pecuniary effects; (2) *in-district factor owners*, who obtain pecuniary gains in production of, as well as benefits in consumption from, $P_j(x)$; (3) *out-of-district factor owners*, who obtain pecuniary gains in production (but no consumption benefits since they do not reside in the local constituency); (4) *in-district consumers who make factor market purchases*,

gains. Nevertheless, the political effect is the same, though now manifested in an aversion to pecuniary losses suffered if demand were to contract (i.e., the flow of projects cease).

[6] There are pecuniary losers as well—namely, other users of project factors who experience rising prices.

who obtain consumption benefits from the project but suffer pecuniary losses in the form of higher prices for factors; and (5) *out-of-district purchasers of factors,* who suffer pecuniary losses through higher prices in factor markets (and who, moreover, obtain no consumption benefits since they do not reside in the district). Public good theorists like Samuelson, Bator, and Head, in their market-failure models of public good provision, typically focus only on type 1 agents and the nonappropriable product of public projects. Cost-benefit studies and welfare analyses (Prest and Turvey 1965; Mishan 1976) also focus on type 1 agents, limiting the role played by the other four types. These studies compute costs and benefits in consumption followed by an appropriate adjustment for pecuniary effects. But these studies may have missed the point. As Aranson and Ordeshook (1978) have emphasized generally, and Stigler (1971) and Peltzman (1976) for the case of regulation, it may well be the appropriable pecuniary gains and losses of factor owners and competing factor users, respectively, that drive the political production process. In this latter view, the consumption benefits of the public project—clean water, mass transit, or whatever—are a by-product of factor owners and factor users seeking pecuniary gains and the avoidance of pecuniary losses, respectively.

How these pecuniary effects are distributed geographically, and whether they are gains or losses, have different kinds of political impacts. Since political representation is geographic, legislators care about who gains and who loses in proportion to their geographic locations. We assume that local gains and losses are politically more significant to the legislator's objectives (reelection and constituency service) than nonlocal effects. Hence representatives use whatever legislative influence they can exercise on behalf of those affected locally by pecuniary effects.

Additionally, there are several reasons to believe that pecuniary gains are exaggerated and pecuniary losses diminished in the representative's political calculus. They relate to the concentration of pecuniary gains and the dispersion of pecuniary losses. First, in what might be termed the "Robert Moses effect" (named after that famous New Yorker who appreciated and exploited it so effectively), is the observation that pecuniary gains in the form of increased jobs, profits, and local tax revenues go to named individuals, firms, and localities from whom the legislator may claim credit and exact tribute. Pecuniary gains may be targeted to constituents; pecuniary losses, on the other hand, are often more widely dispersed, falling on constituents and nonconstituents alike. Second, pecuniary losses, principally through higher prices in factor markets, are not always fully linked to the effects of increased factor demand from the project in

question. Indeed, the illusion may be such that pecuniary losers are unable to distinguish the source of their losses from general price inflation. Hence there is a perceptual asymmetry between pecuniary gains and losses. Accompanying this asymmetry in perception is an asymmetry in capacity to convert perceptions of gains and losses into political influence. Third, then, as Peltzman (1976) has noted in another context, gainers typically are smaller in number, more cohesive in political interest, and, consequently, better organized politically. They are capable of rewarding the local legislator for delivering the bacon in a fashion in which pecuniary losers are unable to punish. The combined impact of the Moses, the illusion, and the Peltzman effects is an exaggerated political importance accorded local pecuniary gainers. The local legislator, then, is strongly encouraged to generate projects with large c_1-type components (vs. c_2-type) and tends to be less concerned with associated pecuniary losses (vs. pecuniary gains).

Since, on the arguments above, there is political value in securing local expenditures for their own sake, the representative, in assessing the project $P_j(x)$, incorporates, on the benefit side of his political calculus, both the consumption benefits his constituents obtain, $b(x)$, and the politically distorted pecuniary effects. Since the latter depend on local expenditures ($c_1[x]$), we write it as $f[c_1(x)]$. For the propositions below, we stipulate $f[c_1(x)] = c_1(x)$ to enable the clear presentation of results unencumbered by mathematical detail; our results are qualitatively unaffected so long as f and f' are positive.[7] Thus, the representative of the single political constituency j seeks a project scale that maximizes his political maximand which, in turn, depends on his constituency's benefits minus its share of the costs: $[b(x) + c_1(x)] - [t_jT(x) + c_3(x)]$.

Model P: A Single Political Constituency

Consider the case of a single national constituency. Then $t_j = 1$ and expenditures earmarked for the constituency now consist of all ex-

[7] A straightforward comparative statics analysis, which we do not pursue here, would examine how alternative specifications of f affect optimal choices. Such an analysis would confirm that as long as f is positive, the qualitative nature of our results stands. As noted, it is not too misleading to equate $f[c_1(x)]$ with $c_1(x)$ or some linear function thereof, namely, $f[c_1(x)] = \beta c_1(x)$, in the relevant range. But only up to a point! If expenditures yield not only consumption benefits but β dollars per dollar expended of additional "benefit," then, subject only to external costs ($c_3[x]$), legislators would be motivated to expend the entire GNP, clearly an absurdity. The function $f[c_1(x)]$ is meant to represent the political (read: electoral) advantage secured by the legislator who delivers $c_1(x)$ dollars of public expenditure to his district. Ultimately this advantage must tail off so that, in a general mathematical analysis, $f[c_1(x)]$ eventually exhibits diminishing returns.

penditures, that is, $c_1(x) + c_2(x)$. With these assumptions and the identity in (1), the political objective function reduces to

$$\begin{aligned}\max_x P(x) &= [b(x) + c_1(x) + c_2(x)] - [c_1(x) + c_2(x) + c_3(x)] \\ &= b(x) - c_3(x).\end{aligned} \tag{4}$$

The first-order condition is

$$b' - c_3' = 0, \tag{5}$$

and the second-order condition requires that

$$b'' - c_3'' < 0.$$

The solution to equation (5) yields point x^P on figure 1. The following theorem shows that $x^P > x^E$.

THEOREM 1: $x^P > x^E$.

PROOF: From assumptions 1 and 2, x^E is the unique global maximum of $E(x) = b(x) - c_1(x) - c_2(x) - c_3(x)$. Consequently, $b'(x^E) - c_1'(x^E) - c_2'(x^E) - c_3'(x^E) = 0$. From assumption 2, $c_1'(x^E) + c_2'(x^E) > 0$; therefore, $P'(x^E) = b'(x^E) - c_3'(x^E) > 0$. But this violates the first-order condition for a maximum of $P(x)$. Indeed, from the properties of strictly concave functions, since $P'(x^E) > 0$, then either (i) $P'(x) = 0$ for $x = x^P > x^E$, or (ii) $P'(x) = 0$ for no finite x. In the first case, x^P is a global interior maximum which exceeds x^E. In the second case, there is no interior maximum since $P(x)$ increases without bound. A fortiori, a project scale larger than x^E is preferred. Q.E.D.

Model N: Non Est e Pluribus Unum, or Every District for Itself

The next stage in our examination of political mechanisms is to partition the single national constituency into multiple, disjoint political units called "districts" with representation in a legislature. Each district, through its representative, is presumed to maximize its net (private) benefits without regard to the costs imposed on other districts. Publicly supported projects are funded through taxes which fall primarily on other districts. Hence, the benefits are concentrated while the costs are diffused.

Any political choice institution consisting of representatives of multiple, disjoint constituencies is characterized by a principle of aggregation—that is, a voting or decision rule—and by the substantive choices made under that decision rule. There are several literatures which address these issues.[8] The theoretical thrust of these

[8] These include research on logrolling, constitutional choice, and distributive policymaking. There is a large literature on each of these topics, but Buchanan and Tullock (1962) remains the best introduction and overview.

literatures is twofold: (1) Why is one institutional decision rule employed rather than another? and (2) What are the policy consequences of a given decision rule? Regarding the first, with special emphasis on majority rule, theorists have focused on the instability of decisive coalitions, the ex ante uncertainty surrounding the composition of winning coalitions, and the cost of organizing and monitoring coalition partners. Their express purpose is to address how institutional actors seek, through suitably arranged institutional practices, to improve their circumstances by evolving coping strategies, formal rules, and other mechanisms (called norms in the sociological literature, rules or binding commitments in the game theory literature, and contracts in the property rights literature). The second question takes institutional practices as fixed and examines the policy decisions implied by those practices. The historical origins of and rationales for institutional rules are of little concern in these latter analyses.

Our chief concern is the second question in which we assess the project choices of a legislature in the distributive policy area. In those policy areas characterized by a project-by-project orientation, the geographic concentration of benefits, and the diffusion of costs, there is abundant evidence that universalism and reciprocity are prevailing decision rules in the U.S. Congress. The former practice assures any interested district a project; the latter, in recognition of the fact that district differences translate into different policy priorities, facilitates a process of mutual support and logrolling. These two practices combine to permit packages of distinct projects earmarked for interested districts to obtain the support even of those without a stake in the package in exchange for reciprocal treatment. Empirical studies, moreover, repeatedly observe the operation of universalistic criteria. Examples include the pork barrel of rivers and harbors (Maass 1951; Ferejohn 1974); model cities and urban renewal (Plott 1968); tax loopholes (Manley 1970); the traditional tariff (Schattschneider 1935); private member bills (Froman 1967); military procurement (Rundquist 1973); and categorical grants-in-aid (Mayhew 1974). Indeed, some scholars observe a tendency to infuse controversial policies with distributive elements in order to build a more inclusive coalition.[9]

Elsewhere (Fiorina 1978; Weingast 1979; Shepsle and Weingast 1981) the conditions under which institutional actors prefer universalistic criteria to pure majority rule are derived. These results, together with the preponderance of empirical evidence cited above, provide the basis for our focus on universalism and its policy consequences. First we explore the consequences of universalism in the

[9] See Stockman (1975) and Fiorina (1978) for illustrations.

extreme in which each legislator independently determines whether to propose a project and, if so, at what scale. The following section then examines the conditions under which there are incentives for legislators to coordinate project selection. Consider legislator j's maximand regarding his district's project:[10]

$$N_j(x) = b(x) + c_1(x) - t_jT(x) - c_3(x). \tag{6}$$

The legislator is presumed to maximize the district's private benefits, $b(x) + c_1(x)$, minus its share of the taxes, $t_jT(x) = t_j[c_1(x) + c_2(x)]$, minus the externalities of the project which fall on the district, $c_3(x)$. This simply reflects the proposition that legislators are reelection oriented and that their prospects are positively associated with the net benefits they deliver to their constituents. According to this reelection mechanism, voter decisions correspond to a "what have you done for me lately?" evaluation.[11]

Maximization of (6) yields the first-order condition

$$b' + c_1' - t_jT' - c_3' = 0,$$

or

$$b' + c_1' = t_j(c_1' + c_2') + c_3'. \tag{7}$$

The second-order condition requires

$$b'' + c_1'' - t_j(c_1'' + c_2'') - c_3'' < 0.$$

Equation (7) has a solution yielding a maximum point x^N for (6) depicted in figure 1. The next theorem establishes that, in a wide set of circumstances, x^N exceeds x^P.

THEOREM 2: If $c_1' > t_j(c_1' + c_2')$, then $x^N > x^P$.

PROOF: Note that $N(x)$ defined in (6) may be recast as

$$N(x) = P(x) + c_1(x) - t_j[c_1(x) + c_2(x)],$$

where $P(x) = b(x) - c_3(x)$. At the maximum of $P(x)$, $x = x^P$,

$$P'(x^P) = b'(x^P) - c_3'(x^P) = 0.$$

This implies that at $x = x^P$

$$\begin{aligned} N'(x^P) &= P'(x^P) + c_1'(x^P) - t_j[c_1'(x^P) + c_2'(x^P)] \\ &= c_1'(x^P) - t_j[c_1'(x^P) + c_2'(x^P)]. \end{aligned}$$

Therefore, the premise of the theorem implies that $N'(x) > 0$ for all x

[10] Eq. (6) does not contain district j's total benefits and costs, only those falling within its control. Theorem 4 below establishes the innocuousness of this omission.

[11] This model of voting is known as retrospective voting in contrast to the prospective voting model, initially popularized by Downs (1957), in which voters respond to promises for future policy.

$\leq x^P$. Consequently, if $N(x)$ possesses a maximum—call it x^N—then $x^N > x^P$. Moreover, if $N(x)$ does not possess a maximum, then our theorem obtains a fortiori since $N(x)$ then increases without bound. Q.E.D.

The condition in the premise of theorem 2, $c_1' > t_j(c_1' + c_2')$, can be presumed to hold in most circumstances. This condition requires that, in district-initiated projects, locally earmarked expenditures grow more rapidly with project scale than local taxes. Since t_j is ordinarily quite small, only a modicum of imagination by legislators is required to find projects with this local expenditure characteristic. For example, if taxes are shared evenly by districts, then $t_j = 1/n$, where n is the number of districts, and the condition becomes

$$c_1' > \frac{1}{n}(c_1' + c_2').$$

That is, so long as district-targeted expenditures grow with project scale at a rate at least $1/n$th as fast as total expenditures, the premise (and hence the conclusion) of theorem 2 holds; in even moderate-sized legislatures this is a fairly weak condition.

More generally, suppose $t_j = t_j(z)$ where z might be any characteristic (examples of which are given below). Then the following comparative statics result may be established.

THEOREM 3: Let x^N be the optimum for (6). Then $dx^N/dz > 0$ if and only if $t_j' < 0$.

PROOF: Rewriting (6), we have

$$N(x) = b(x) + [1 - t_j(z)]c_1(x) - t_j(z)c_2(x) - c_3(x),$$

and first-order condition

$$N'(x) = b'(x) + [1 - t_j(z)]c_1'(x) - t_j(z)c_2'(x) - c_3'(x) = 0.$$

The second-order requirement is

$$N''(x) = b''(x) + [1 - t_j(z)]c_1''(x) - t_j(z)c_2''(x) - c_3''(x) < 0.$$

Totally differentiating the first-order condition yields

$$\{b''(x) + [1 - t_j(z)]c_1''(x) - t_j(z)c_2''(x) - c_3''(x)\}dx$$
$$= [c_1'(x) + c_2'(x)]t_j'(z)dz.$$

Thus,

$$\frac{dx}{dz} = \frac{[c_1'(x) + c_2'(x)]}{\{b''(x) + [1 - t_j(z)]c_1''(x) - t_j(z)c_2''(x) - c_3''(x)\}}t_j'(z).$$

The numerator of the coefficient on the right-hand side is positive (from assumption 2) and the denominator is negative (from the

second-order condition). The coefficient, therefore, is negative, establishing that the sign of dx/dz is opposite that of $t_j'(z)$:

$$t_j'(z) < 0 \rightarrow \frac{dx}{dz} > 0,$$

$$t_j'(z) > 0 \rightarrow \frac{dx}{dz} < 0.$$

Q.E.D.

COROLLARY (The Law of $1/n$): If district tax share is a declining function of the number of districts (n), then the degree of inefficiency in project scale ($x^N - x^E$) is an increasing function of the number of districts.

PROOF: Let $t_j = t_j(n)$ with $t_j'(n) < 0$. Then, from theorem 3, $dx^N/dn > 0$. Q.E.D.

The corollary indicates that when taxes are apportioned as a decreasing function of the number of political units—for example, $t_j(n) = 1/n$ for all j—then the optimum project scale for any district grows as the polity is more finely partitioned into districts.[12] Theorem 3, however, is more general, for it applies to tax mechanisms that may be the function of any politically relevant characteristic. If a district's tax share is a decreasing function of certain of its legislator's institutional characteristics (membership or influence on tax-writing committee), political characteristics of its representative (is he a committee chairman? is he associated with the majority party?), or economic or demographic characteristics of the district (proportion of families below poverty level, proportion of population above age 65), then we can associate increasingly inefficient projects with particular kinds of districts as defined by these tax-relevant characteristics. In all these cases, theorem 3 indicates that the equilibrium scale of a district's project (and, given the assumptions, its degree of inefficiency) changes with respect to some tax criterion in precisely the opposite way the tax share changes with respect to that criterion.

Institutional Incentives to Restrict Project Size

We have just shown that decentralized choice by a representative legislature characterized by a universalism mechanism, a tax-sharing

[12] This analysis presumes that the only change in $N(x)$ following a change in the number of districts is the tax rate, t_j. If, however, $c_1(x)$, $c_2(x)$, and $c_3(x)$ depend on the configuration of districts, then two countervailing tendencies may be observed. Increasing the number of districts (1) transforms some portion of $c_1(x)$ into $c_2(x)$ and (2) decreases the tax share of the district. Since the first effect reduces and the second increases the optimal project scale, the net effect is ambiguous without further specification.

rule, and a political objective function yields a vector of projects $x^N = (x_1^N, \ldots, x_n^N)$. Theorems 1 and 2 provide the conditions in which $x_i^N > x_i^E$, $i = 1, \ldots, n$. However, x_i^N, the project scale for district i that maximizes its legislator's political objective function (6), is computed in isolation of computations by other legislators; therefore, it does not take the expenditure and tax externalities (positive and negative) generated by those other projects into account.

In this section, we seek to determine whether or not there is an institutional basis for restraining unbridled universalism.[13] In particular, we turn to an examination of packages of projects and seek to discover whether there exists a package of constitutionally restricted projects, $x^C = (x_1^C, \ldots, x_n^C)$, with the property that $x^C >_i x^N$ (where $>_i$ is the preference order of district i over packages of projects). To accomplish this, consider the complete political maximand of legislator j, $B^j(x_1, \ldots, x_n)$, which rewrites (6) to incorporate the effects of projects in other districts. The net benefits to district j consist of b-type benefits from its own project, $b(x_j)$; c_1-type expenditures from its own project, $c_1(x_j)$; c_2-type expenditures spent in district j from other projects, $\Sigma_{i \neq j} c_{2j}(x_i)$; nonexpenditure costs from its own project $c_3(x_j)$; and its tax share of the total expenditures, $t_j \Sigma_{i=1}^n [c_1(x_i) + c_2(x_i)]$. Thus

$$B^j(x_1, \ldots, x_n) = \left[b(x_j) + c_1(x_j) + \sum_{i \neq j} c_{2j}(x_i)\right] - \left\{c_3(x_j) + t_j \sum_{i=1}^{n} [c_1(x_i) + c_2(x_i)]\right\}. \tag{8}$$

Before considering the possibility of restrictions on project size, we first characterize choice under the complete political maximand, $B^j(x_1, \ldots, x_n)$. Calling x_j^B legislator j's solution to the maximization of (8), we have:

THEOREM 4: $x_j^B = x_j^N$.

PROOF: Equation (8) may be rewritten

$$B^j = \{b(x_j) + c_1(x_j) - c_3(x_j) - t_j[c_1(x_j) + c_2(x_j)]\} + \left\{\sum_{i \neq j} c_{2j}(x_i) - t_j \sum_{i \neq j} [c_1(x_i) + c_2(x_i)]\right\}, \tag{9}$$

which, from (6), becomes

$$B^j = N_j(x_j) + \left\{\sum_{i \neq j} c_{2j}(x_i) - t_j \sum_{i \neq j} [c_1(x_i) + c_2(x_i)]\right\}. \tag{10}$$

[13] In Shepsle and Weingast (1981), we describe results that demonstrate the ex ante superiority of universalism to pure majority rule in the eyes of each legislator. Here we ask whether a restricted form of universalism, in turn, is superior to pure universalism.

Since the terms in braces are not functions of x_j, the first- and second-order conditions for B^j are the same as those for $N_j(x_j)$; thus, their respective maxima are the same. Q.E.D.

Theorem 4 establishes that our initial focus on N_j (instead of B^j) involved no loss of generality for questions about decentralized project choice. Moreover, theorem 4 establishes that x_j^N is a Nash strategy for legislator j and that the project vector, x^N, is a Nash equilibrium.[14]

Turning to the role of cooperatively imposed restrictions on project size, we state a characterization theorem that establishes the circumstances under which districts (or their representatives) prefer a restricted collection of projects, x^C, to the collection under unrestrained universalism, x^N. For the vector of projects $x^K = (x_1^K, \ldots, x_n^K)$, let $E_j(x^K)$ represent the tax and expenditure effects of other projects on district j. Specifically,

$$E_j(x^K) = \sum_{l \neq j} \{t_j[c_1(x_l^K) + c_2(x_l^K)] - c_{2j}(x_l^K)\}.$$

In effect, $E_j(x^K)$ is the tax bill for district j for all other projects net of local expenditures to district j from these other projects—that is, net negative external costs.

THEOREM 5: $x^C >_j x^N$ if and only if $N_j(x^N) - N_j(x^C) < E_j(x^N) - E_j(x^C)$.

Theorem 5 conveys the following idea: Legislator j has an interest in substituting a collection of projects $x^C = (x_1^C, \ldots, x_n^C)$ for $x^N = (x_1^N, \ldots, x_n^N)$ if and only if his district's reduction in political benefits from the reduced scale of its own project is compensated for by a concomitant reduction in its burden of net negative external costs. There are several ways in which this may fail to hold for a given district j. For example, (i) district j has a sufficiently small tax share, t_j; (ii) district j is, disproportionately, a source of project inputs for other districts, in which case the $c_{2j}(x_l)$ terms are large; and (iii) the political benefits for district j with project scale x_j^N, $b(x_j^N) + c_1(x_j^N)$, are inordinately large. In each of these cases, the condition may fail, implying the absence of a unanimous preference for x^C over x^N. Moreover, there do not appear to be any interesting properties associated with the family of functions $b(x)$, $c(x)$, and $t(x)$ that satisfy the condition in theorem 5; nor is there any ex ante basis for supposing that the condition in theorem 5 will be encountered in empirical settings.

Despite the lack of unanimous preference for restriction under all

[14] This view interprets our model of unrestricted universalism as a game. The choice set for player j is the size of his project, x_j, when project choice is governed by a universalism mechanism. The Nash equilibrium follows from the separability of the positive and negative externalities of other projects in (10). As a result, each district makes its own decisions without attending to the externalities it produces or consumes.

circumstances, legislators have something to gain by properly accounting for the pecuniary externalities of project scale selection. The next theorem provides some insight into the optimal set of projects which internalizes these effects.

THEOREM 6: x^P maximizes $\Sigma_{j=1}^{n} B^j(x)$, given in (8).

PROOF:

$$L = \sum_{j=1}^{n} B^j(x) = \sum_{j=1}^{n} \Big\{ b(x_j) + c_1(x_j) + \sum_{i \neq j} c_{2j}(x_i) - c_3(x_j) - t_j \sum_{i=1}^{n} [c_1(x_i) + c_2(x_i)] \Big\}$$

$$= \sum_{j=1}^{n} [b(x_j) + c_1(x_j) + c_2(x_j) - c_3(x_j) - c_1(x_j) - c_2(x_j)]$$

$$= \sum_{j=1}^{n} [b(x_j) - c_3(x_j)].$$

The first-order conditions are

$$\frac{\partial L}{\partial x_j} = b'(x_j) - c_3'(x_j) = 0, \qquad j = 1, \ldots, n.$$

Noting that $(\partial^2 L)/(\partial x_i \partial x_j) = 0$ for $i \neq j$, the second-order conditions are

$$\frac{\partial^2 L}{\partial x_j^2} = b''(x_j) - c_3''(x_j) < 0, \qquad j = 1, \ldots, n.$$

These equations and inequalities are the same as (5), which identifies the vector x^P in figure 1. Q.E.D.

Theorem 6 shows that total political net benefits are maximized with the vector $x^P = (x_1^P, \ldots, x_n^P)$. Potential gains may be captured if the system of universalistic project selection is amended so that, while each district is assured a project, the project scale is determined as if there were but a single district (as in model P above). Notice that theorem 6 does not assert that $x^P \succ_j x^N, j = 1, \ldots, n$ (substituting x^P for x^C, theorem 5 shows the restricted circumstances in which this will hold). What theorem 6 does assert is that a compensation scheme which redistributes net benefits is feasible so that x^P together with this compensation is preferred by all districts not only to x^N but to any other omnibus of projects.

It is occasionally asserted that the distributive politics game is a prisoner's dilemma in the economic sense—that unrestrained universalism produces a project package that is an economically inefficient Nash equilibrium, on the one hand, and is unanimously regarded as

less preferable than x^E ($x^E \succ_j x^N, j = 1, \ldots, n$) on the other. Theorems 1, 2, and 4 establish the first part of this assertion. However, retaining the political conceptualization of net benefits given in (8), theorem 5 shows that the second part of the assertion does not always follow. The collection x^E is not always unanimously regarded as preferable to x^N. More importantly, theorem 6 shows that even when all representatives favor restrictions, the *politically* optimal set of projects is x^P, not x^E.

This discussion underscores the basic point of the paper. The efficient collection, x^E, though normatively attractive in welfare analysis, is not always behaviorally relevant. The implication of the political maximand (8) and theorems 1, 2, 5, and 6 is that legislators hold no brief for efficiency, per se, either with regard to their own project selection or a package of such projects.

Extensions and Discussion

The model developed in the previous sections roots the inefficiency of distributive politics in democratic mechanisms and especially in the geographic basis of political constituencies. This latter feature produces two independent sources of bias. First, locally targeted expenditures are counted by the local constituency as benefits. Second, the districting mechanism in conjunction with the taxation system provides incentives to increase project size beyond the efficient point by attenuating the relationship between beneficiaries and revenue sources. A cooperative legislature has no incentive to remove entirely these sources of inefficiency (beyond that described in the discussion following theorem 6).

In this section, we examine several related themes and applications. Each of these is either an extension of our model to domains beyond that of traditional distributive policy or a specialization of our model to substantively relevant cases of distributive policy.

Congressional Limitation on Project Size

A well-known behavioral mechanism has operated in Congress since the 1880s that restrains the attempts of legislators to fund their pet projects. At the authorization stage, a universalism mechanism is at work—the annual omnibus public works bill, for example, typically contains authorizations for projects in most congressional districts. Following the passage of authorizing legislation comes the separate stage of actually appropriating monies. Here, the Committee on Appropriations systematically scales down each project (for a description of this process, see Fenno [1966] and Ferejohn [1974]).

Our model sheds some light on this well-established congressional practice. First, suppose the premise of theorem 5 holds, where $x^P = x^C$. Hence, $x^P >_j x^N, j = 1, \ldots, n$. Second, assume legislator j and only legislator j knows x_j^P and x_j^N. Theorem 5 implies that legislators unanimously favor a mechanism reducing project sizes from the uncoordinated choice, x^N. But there does not appear to be a straightforward "demand-revealing" mechanism inducing legislators to announce anything other than their maximizing Nash strategy, x_j^N. Thus, in the face of imperfect information, the simpler mechanism of scaling down all projects—the current practice of the appropriations committee—may yield a vector of project sizes less than x^N and preferred by all legislators.[15]

Examination of Interesting Subclasses

We mention briefly, as an indication of the utility of this approach, some interesting special cases of the political maximand (6):

$$N(x_j) = b(x_j) + c_1(x_j) - c_3(x_j) - t_j[c_1(x_j) + c_2(x_j)].$$

The details may be provided by the reader.

1. *Benefit tax.*—This tax scheme requires the district to pay the entire cost of a strictly local public project ($t_j = 1$). The maximand becomes

$$N_j(x_j) = b(x_j) - c_2(x_j) - c_3(x_j),$$

and assumptions 1 and 2 imply a project scale x_j^* with the property $x_j^E \leq x_j^* \leq x_j^P < x_j^N$. Note that if $c_2(x_j) = 0$, then $x_j^* = x_j^P$ whereas, if $c_1(x_j) = 0$, then $x_j^* = x_j^E$. This last fact suggests that a benefit tax in conjunction with no local expenditures is a sufficient condition for public sector efficiency.

2. *Free-ride tax.*—In this case, $t_j = 0$ and the maximand is

$$N(x_j) = b(x_j) + c_1(x_j) - c_3(x_j),$$

which implies a scale $x_j^* > x_j^N$. This case approximates local public goods, financed by user fees, where the local residents rarely number among the users. For example, if Yellowstone National Park were financed by user fees assessed states in proportion to their respective

[15] A complete analysis must resolve two issues. First, as a consequence of the scaling down practice by the Appropriations Committee, will the legislator strategically seek a project scale in excess of x_j^N so that, when it is scaled down, it will eventually reach size x_j^N? Alternatively, are there sanctions discouraging artificial inflation of project scale at the authorizing stage? Second, are there sanctions available to be applied against the Appropriations Committee to ensure it does not scale down too much? See Fenno (1966) for some empirical details. These issues concerning demand revelation mechanisms are theoretically intriguing but take us too far afield to be dealt with here.

share of the population using the park, then Wyoming residents receive a cheap, if not free, ride on taxes.

3. *Foreign aid versus military assistance.*—We can interpret the political popularity of the latter and unpopularity of the former via the following calculation. Let x^F and x^M describe levels of the two forms of aid; assume $b(x_j^F) = b(x_j^M)$, ceteris paribus, for any political constituency (assumed, in any event, to be small unless the constituency contains partisans of or emigrants from the benefiting country); for some constituencies (producers of military hardware), $c_1(x_j^M) > 0$; for all constituencies, on the other hand, $c_1(x_j^F) = 0$; for all constituencies, $c_3(x_j^M) = c_3(x_j^F) = 0$. Thus,

$$N(x_j^M) = b(x_j^M) + (1 - t_j)c_1(x_j^M) - t_j c_2(x_j^M),$$

$$N(x_j^F) = b(x_j^F) - t_j c_2(x_j^F).$$

Assumptions 1 and 2 imply $(x_j^M)^* > (x_j^F)^*$, an observation consistent with their respective popularities. It also makes sense of the recent policy innovation of attaching strings to foreign aid requiring recipients to make purchases in the United States (c_1-type expenditures).

4. *Rube Goldberg machines and military bases.*—Suppose $b(x_j) = 0$ for a project in a given constituency, j. Then

$$N(x_j) = (1 - t_j)c_1(x_j) - t_j c_2(x_j) - c_3(x_j).$$

Even though $b(x_j) = 0$, the (politically) optimal project scale may be greater than zero. Specifically,

$$x_j^N > 0 \text{ if } N'(0) > 0, \text{ or } (1 - t_j)c_1'(0) - t_j c_2'(0) - c_3'(0) > 0.$$

Consider the case of many military bases (and other Rube Goldberg machines) which, by the Department of Defense's own admission, provide virtually no contribution to defense ($b[x] = 0$). These nevertheless remain attractive to local constituencies (hence $x_j^N > 0$) because of the overriding importance of these projects to the local economy in the form of c_1-type benefits. Throughout this paper, we have emphasized the political inappropriateness of economic net benefits as a relevant decision criterion. In this special case, it appears that even the absence of economic benefits altogether is not a disqualifying characteristic in political choice.

Generalization to Non–Pork Barrel Policies[16]

One of the central features of our models is the unpacking of costs in politically relevant ways in which we distinguish project costs returned

[16] These themes are developed in more detail in Shepsle and Weingast (1980).

to the district as expenditures, costs returned to other districts, and nonexpenditure costs borne within the district. We focused, however, on distributive or pork barrel projects, defined as projects whose benefits are geographically concentrated and whose costs are spread through general taxation. Two extensions offer further insight into nondistributive policies. The first distinguishes an additional cost, $c_4(x)$, or external, nonexpenditure costs borne by other districts. The second defines another source of economic benefits, $b_2(x)$, that accrues to other constituencies as a consequence of a project in a particular district.

1. *Additional external costs.*—If additional external costs of the jth project spill over into other districts, it can easily be shown that the degree of inefficiency increases when governed by decentralized project choice under the political maximand, $N(x)$. Since the model now divides externalities into their politically relevant components, $c_3(x)$ and $c_4(x)$, we can make further observations about the degree to which the political system can be relied upon to internalize externalities associated with public activities. If the externalities are not too large (in the sense that they are local and do not extend into neighboring districts), the public sector action governed by a representative legislature internalizes them. However, if they are large, public sector action may not.

The politically relevant (though economically arbitrary) distinction implied by district boundaries suggests that the public sector can only provide certain categories of public goods which are not available through private market arrangements. Hence, jurisdictional questions become of paramount importance when producing a local public good like flood walls along a river which divides two political jurisdictions. Some of the most infamous cases of pork barrel politics illustrate this point: The flood walls along the lower part of the Mississippi River, which divides Louisiana from Mississippi, are 3 feet higher on the Mississippi side (see Ferejohn 1974, pp. 56–58). Similarly, the levees on the Indiana side of the Wabash River are higher than those on the Illinois side. Thus, a universalistic representative legislature is biased toward projects with low c_3-type costs while failing to consider c_4-type costs.

2. *Additional external benefits.*—Let $b(x_j) = b_1(x_j) + b_2(x_j)$ where $b_1(x_j)$ are the benefits of the jth project concentrated in district j and $b_2(x_j)$ are the benefits consumed by residents of other districts (presumed zero throughout the body of the paper). That is, $b_2(x_j)$ is the positive consumption externalities (as compared with $c_2[x_j]$, which are the positive production externalities).

Because the benefits outside the district are not readily internalized

under a distributive policy mechanism, large multidistrict (multistate) projects are likely to be rare relative to projects with concentrated local benefits. Consequently, multidistrict public goods (in the economic sense) are not only likely to be underproduced by a market mechanism but by a representative legislature as well (see Aranson and Ordeshook 1978). A universalistic representative legislature is biased toward projects with high b_1-type benefits, while failing to internalize b_2-type benefits. Thus, both positive and negative externalities adversely affect public as well as private provision.

Conclusion

Throughout this paper we have focused on the sources of inefficiency in public decision making. Our model demonstrates that democratic institutions play an important role. Three mechanisms were shown to influence the politically optimal project choice: the political cost-accounting mechanism, the districting mechanism, and the taxation mechanism. These features of the political economy systematically transform the economic benefits and costs into political counterparts. Since it is the latter that determine the maximands for political actors and not their economic counterparts, these govern political choice.

While our modeling of the districting and taxation mechanisms is straightforward and uncontroversial, there are circumstances in which our treatment of the incidence of gains and losses from local expenditures is implausible. We have presumed that pecuniary gainers figure more prominently than pecuniary losers in a legislator's reelection constituency. However, under some circumstances this may not be true. A legislator's reelection constituency, for example, may be dominated not by factor owners of a public project but by those who would bear the brunt of the pecuniary losses and the nonpecuniary external costs of the project. We would not, therefore, expect the legislator to seek such projects. Indeed, since the menu of distributive programs is sufficiently diverse, there normally is something available for everyone. Thus we tend to find reclamation projects in the West, locks and dams in river districts with an active construction industry, and wildlife refuges in Sierra Club districts. Because of this diversity in policy preferences and program categories, the logic supporting the political distortion of pecuniary incidences continues to hold.

Our principal conclusion is that since political institutions fundamentally alter the perceptions and incidence of benefits and costs, they systematically bias project choices away from the efficient outcomes. In the context of distributive politics, this was shown to imply larger projects and programs than are economically warranted.

References

Aranson, Peter H., and Ordeshook, Peter C. "The Political Bases of Public Sector Growth in a Representative Democracy." Paper presented at the Conference on the Causes and Consequences of Public Sector Growth, Dorado Beach, Puerto Rico, October 1978.

Breton, Albert. *The Economic Theory of Representative Government.* Chicago: Aldine, 1974.

Buchanan, James M., and Tullock, Gordon. *The Calculus of Consent: Logical Foundations of Constitutional Democracy.* Ann Arbor: Univ. Michigan Press, 1962.

Downs, Anthony. *An Economic Theory of Democracy.* New York: Harper & Row, 1957.

———. *Inside Bureaucracy.* Boston: Little, Brown, 1967.

Fenno, Richard F., Jr. *The Power of the Purse: Appropriations Politics in Congress.* Boston: Little, Brown, 1966.

Ferejohn, John A. *Pork Barrel Politics: Rivers and Harbors Legislation, 1947–68.* Stanford, Calif.: Stanford Univ. Press, 1974.

Ferejohn, John A., and Fiorina, Morris P. "Purposive Models of Legislative Behavior." *A.E.R. Papers and Proc.* 65 (May 1975): 407–14.

Fiorina, Morris P. "Legislative Facilitation of Government Growth: Universalism and Reciprocity Practices in Majority Rule Institutions." Paper presented at the Conference on the Causes and Consequences of Public Sector Growth, Dorado Beach, Puerto Rico, October 1978.

Fiorina, Morris P., and Noll, Roger G. "Voters, Bureaucrats and Legislators: A Rational Choice Perspective on the Growth of Bureaucracy." *J. Public Econ.* 9 (April 1978): 239–54.

Froman, Lewis A., Jr. *The Congressional Process: Strategies, Rules, and Procedures.* Boston: Little, Brown, 1967.

Lowi, Theodore J. "American Business, Public Policy, Case-Studies, and Political Theory." *World Politics* 16 (July 1964): 677–715.

———. *The End of Liberalism: The Second Republic of the United States.* 2d ed. New York: Norton, 1979.

Maass, Arthur. *Muddy Waters: The Army Engineers and the Nation's Rivers.* Cambridge, Mass.: Harvard Univ. Press, 1951.

McKean, Roland N. *Efficiency in Government through Systems Analysis, with Emphasis on Water Resources Development.* New York: Wiley, 1958.

Manley, John F. *The Politics of Finance: The House Committee on Ways and Means.* Boston: Little, Brown, 1970.

Mayhew, David. *Congress: The Electoral Connection.* New Haven, Conn.: Yale Univ. Press, 1974.

Mishan, E. J. *Cost-Benefit Analysis.* 2d ed. New York: Praeger, 1976.

Niskanen, William A. *Bureaucracy and Representative Government.* Chicago: Aldine-Atherton, 1971.

Peltzman, Sam. "Toward a More General Theory of Regulation." *J. Law and Econ.* 19 (August 1976): 211–40.

Plott, Charles R. "Some Organizational Influences on Urban Renewal Decisions." *A.E.R. Papers and Proc.* 58 (May 1968): 306–21.

Prest, Alan R., and Turvey, Ralph. "Cost-Benefit Analysis: A Survey." *Econ. J.* 75 (December 1965): 683–735.

Riker, William H., and Brams, Steven J. "The Paradox of Vote Trading." *American Polit. Sci. Rev.* 67 (December 1973): 1235–47.

under a distributive policy mechanism, large multidistrict (multistate) projects are likely to be rare relative to projects with concentrated local benefits. Consequently, multidistrict public goods (in the economic sense) are not only likely to be underproduced by a market mechanism but by a representative legislature as well (see Aranson and Ordeshook 1978). A universalistic representative legislature is biased toward projects with high b_1-type benefits, while failing to internalize b_2-type benefits. Thus, both positive and negative externalities adversely affect public as well as private provision.

Conclusion

Throughout this paper we have focused on the sources of inefficiency in public decision making. Our model demonstrates that democratic institutions play an important role. Three mechanisms were shown to influence the politically optimal project choice: the political cost-accounting mechanism, the districting mechanism, and the taxation mechanism. These features of the political economy systematically transform the economic benefits and costs into political counterparts. Since it is the latter that determine the maximands for political actors and not their economic counterparts, these govern political choice.

While our modeling of the districting and taxation mechanisms is straightforward and uncontroversial, there are circumstances in which our treatment of the incidence of gains and losses from local expenditures is implausible. We have presumed that pecuniary gainers figure more prominently than pecuniary losers in a legislator's reelection constituency. However, under some circumstances this may not be true. A legislator's reelection constituency, for example, may be dominated not by factor owners of a public project but by those who would bear the brunt of the pecuniary losses and the nonpecuniary external costs of the project. We would not, therefore, expect the legislator to seek such projects. Indeed, since the menu of distributive programs is sufficiently diverse, there normally is something available for everyone. Thus we tend to find reclamation projects in the West, locks and dams in river districts with an active construction industry, and wildlife refuges in Sierra Club districts. Because of this diversity in policy preferences and program categories, the logic supporting the political distortion of pecuniary incidences continues to hold.

Our principal conclusion is that since political institutions fundamentally alter the perceptions and incidence of benefits and costs, they systematically bias project choices away from the efficient outcomes. In the context of distributive politics, this was shown to imply larger projects and programs than are economically warranted.

References

Aranson, Peter H., and Ordeshook, Peter C. "The Political Bases of Public Sector Growth in a Representative Democracy." Paper presented at the Conference on the Causes and Consequences of Public Sector Growth, Dorado Beach, Puerto Rico, October 1978.

Breton, Albert. *The Economic Theory of Representative Government*. Chicago: Aldine, 1974.

Buchanan, James M., and Tullock, Gordon. *The Calculus of Consent: Logical Foundations of Constitutional Democracy*. Ann Arbor. Univ. Michigan Press, 1962.

Downs, Anthony. *An Economic Theory of Democracy*. New York: Harper & Row, 1957.

———. *Inside Bureaucracy*. Boston: Little, Brown, 1967.

Fenno, Richard F., Jr. *The Power of the Purse: Appropriations Politics in Congress*. Boston: Little, Brown, 1966.

Ferejohn, John A. *Pork Barrel Politics: Rivers and Harbors Legislation, 1947–68*. Stanford, Calif.: Stanford Univ. Press, 1974.

Ferejohn, John A., and Fiorina, Morris P. "Purposive Models of Legislative Behavior." *A.E.R. Papers and Proc.* 65 (May 1975): 407–14.

Fiorina, Morris P. "Legislative Facilitation of Government Growth: Universalism and Reciprocity Practices in Majority Rule Institutions." Paper presented at the Conference on the Causes and Consequences of Public Sector Growth, Dorado Beach, Puerto Rico, October 1978.

Fiorina, Morris P., and Noll, Roger G. "Voters, Bureaucrats and Legislators: A Rational Choice Perspective on the Growth of Bureaucracy." *J. Public Econ.* 9 (April 1978): 239–54.

Froman, Lewis A., Jr. *The Congressional Process: Strategies, Rules, and Procedures*. Boston: Little, Brown, 1967.

Lowi, Theodore J. "American Business, Public Policy, Case-Studies, and Political Theory." *World Politics* 16 (July 1964): 677–715.

———. *The End of Liberalism: The Second Republic of the United States*. 2d ed. New York: Norton, 1979.

Maass, Arthur. *Muddy Waters: The Army Engineers and the Nation's Rivers*. Cambridge, Mass.: Harvard Univ. Press, 1951.

McKean, Roland N. *Efficiency in Government through Systems Analysis, with Emphasis on Water Resources Development*. New York: Wiley, 1958.

Manley, John F. *The Politics of Finance: The House Committee on Ways and Means*. Boston: Little, Brown, 1970.

Mayhew, David. *Congress: The Electoral Connection*. New Haven, Conn.: Yale Univ. Press, 1974.

Mishan, E. J. *Cost-Benefit Analysis*. 2d ed. New York: Praeger, 1976.

Niskanen, William A. *Bureaucracy and Representative Government*. Chicago: Aldine-Atherton, 1971.

Peltzman, Sam. "Toward a More General Theory of Regulation." *J. Law and Econ.* 19 (August 1976): 211–40.

Plott, Charles R. "Some Organizational Influences on Urban Renewal Decisions." *A.E.R. Papers and Proc.* 58 (May 1968): 306–21.

Prest, Alan R., and Turvey, Ralph. "Cost-Benefit Analysis: A Survey." *Econ. J.* 75 (December 1965): 683–735.

Riker, William H., and Brams, Steven J. "The Paradox of Vote Trading." *American Polit. Sci. Rev.* 67 (December 1973): 1235–47.

Rundquist, Barry S. "Congressional Influences on the Distribution of Prime Military Contracts." Ph.D. dissertation, Stanford Univ., 1973.

Schattschneider, Elmer E. *Politics, Pressures and the Tariff.* Englewood Cliffs, N.J.: Prentice-Hall, 1935.

Shepsle, Kenneth A., and Weingast, Barry R. "Political Solutions to Market Problems: The Political Incidence of Economic Benefits and Costs." Paper presented at the annual meeting of the Public Choice Society, San Francisco, March 1980.

———. "Political Preferences for the Pork Barrel: A Generalization." *American J. Polit. Sci.* 25 (February 1981): 96–111.

Stigler, George J. "The Theory of Economic Regulation." *Bell J. Econ. and Management Sci.* 2 (Spring 1971): 3–21.

Stockman, David. "The Social Pork Barrel." *Public Interest* 39 (Spring 1975): 3–30.

Wagner, Richard E. "Revenue Structure, Fiscal Illusion, and Budgetary Choice." *Public Choice* 25 (Spring 1976): 45–61.

Weingast, Barry R. "A Rational Choice Perspective on Congressional Norms." *American J. Polit. Sci.* 23 (May 1979): 245–63.

Flexible Exchange Rates, Prices, and the Role of "News": Lessons from the 1970s

Jacob A. Frenkel

University of Chicago and National Bureau of Economic Research

This paper analyzes the key issues and lessons from the experience with flexible exchange rates during the 1970s. It analyzes the efficiency of the foreign-exchange market and the volatility of exchange rates, as well as the relationships between exchange rates and interest rates. A key distinction is made between anticipated and unanticipated events, and it is shown that the key factor affecting exchange rates has been "news." The analysis then proceeds to analyze the relationship between exchange rates and prices. The deviations from purchasing power parities are being interpreted in terms of the modern asset-market approach to the exchange rate.

I. Introduction

Recent experience with flexible exchange-rate systems has led to renewed interest in the operation of foreign-exchange markets as

An earlier version of this paper was prepared for a conference (Stabilization Policy: Lessons from the 1970's and Implications for the 1980's) sponsored by the Federal Reserve Bank of St. Louis and the Center for the Study of American Business at Washington University held at the Federal Reserve Bank of St. Louis on October 19–20, 1979. I am indebted to Lauren J. Feinstone for helpful suggestions and efficient research assistance and to the National Science Foundation (grant SOC 78-14480) for financial support. I have benefited from useful comments by Andrew Abel; William Branson; Sebastian Edwards; Stanley Fischer; Craig S. Hakkio; Edi Karni; Paul Krugman; Leonardo Leiderman; Robert E. Lucas, Jr.; Allan Meltzer; Michael L. Mussa; Sam Peltzman; Nasser Saidi; and Roland Vaubel, as well as of participants in seminars held at the NBER, Columbia University, New York University, Harvard University, the University of Virginia, the University of Rochester, Oxford University, Tel-Aviv University, and the International Monetary Fund. This research is part of the NBER's program in international studies. The views expressed are those of the author and not necessarily those of the NBER.

Journal of Political Economy, 1981, vol. 89, no. 4]
© 1981 by The University of Chicago 0022-3808/81/8904-0003$01.50

reflected in many recent studies of the principal determinants of exchange rates. The 1970s witnessed the dramatic alteration of the international monetary system from a regime of pegged exchange rates which prevailed for about a quarter of a century (since the Bretton Woods Conference) into a regime of flexible (though managed) rates. As a consequence of the emergence of the new legal and economic system, traders, national governments, and international organizations were confronted with new economic problems, choices, and instruments. During the 1970s exchange rates fluctuated widely and inflation rates accelerated. The international monetary system had to accommodate extraordinarily large oil-related shocks which affected trade flows in goods and assets. Huge oil payments had to be recycled. Uncertainties concerning future developments in international politics reached new heights, and the prospects for the world economy got gloomier. These developments placed unprecedented pressures on the markets for foreign exchange as well as on other asset markets. They were associated with a large slide in the value of the U.S. dollar and resulted in speeding up the creation of new institutions like the European Monetary System which provided the formal framework for the management of exchange rates among members. The increased interdependence among countries and the realization that exchange-rate policies by one national government exert influence on other economies have also induced legal responses from international organizations. For example, in late April 1977, the executive board of the International Monetary Fund approved the details of the second amendment to article IV of the amended Articles of Agreement dealing with the principles and procedures for surveillance of exchange-rate policies of member countries.

These developments provide the background for this paper which is intended to sum up the relevant evidence bearing on a set of related questions and to present a brief survey of key issues and lessons from the experience with floating rates during the 1970s. The main orientation of the paper is empirical and the analysis is based on the experience of three exchange rates involving the dollar/pound, the dollar/French franc and the dollar/deutsche mark (DM). Section II provides an analysis of the efficiency of foreign-exchange markets by examining the relationship between spot and forward exchange rates. The extent of exchange-rate volatility is also examined. This analysis of the foreign-exchange markets sheds light on several questions including (i) whether exchange rates fluctuate "excessively"; (ii) whether speculation in the foreign-exchange markets is destabilizing; (iii) whether there is "insufficient" speculation in the foreign-exchange markets; and (iv) whether there is evidence for market failure in the sense that there are unexploited profit opportunities.

These issues are significant in assessing the performance of floating rates as well as in evaluating the need for government intervention in the foreign-exchange markets. The analytical framework that is used for interpreting the volatility of exchange rates and the association between spot and forward rates is the modern theory of exchange-rate determination. Within this framework, exchange rates are viewed as the prices of assets that are traded in organized markets and, like the prices of other assets, are strongly influenced by expectations about future events.

The relationship between exchange rates and interest rates is analyzed in Section III. One of the key issues that is raised in this section is the distinction between anticipated and unanticipated changes in rates of interest. As an analytical matter this distinction is important, because the modern approach to exchange-rate determination implies that exchange rates are strongly influenced by "news," which by definition is unpredictable. Therefore, it is unanticipated rather than anticipated changes in interest rates that should be closely associated with changes in exchange rates. This prediction is tested empirically.

Section IV analyzes the relationship between exchange rates and prices by examining the patterns of deviations from purchasing power parities. The main point that is being emphasized is that there is an important intrinsic difference between exchange rates and national price levels. Exchange rates are more sensitive to expectations concerning future events than national price levels. As a result, in periods which are dominated by "news" which alters expectations, exchange rates are likely to be more volatile, and departures from purchasing power parities are likely to be the rule rather than the exception. The analysis of the relationship between exchange rates and prices is relevant for assessing whether the flexible exchange-rate system was successful in providing national economies with an added degree of insulation from foreign shocks and whether it provided policy makers with an added instrument for the conduct of macroeconomic policy. The evidence regarding deviations from purchasing power parities is also relevant for determining whether there is a case for managed float. Section V contains concluding remarks.

II. The Efficiency of the Foreign-Exchange Market and the Movement of Exchange Rates

In this section I analyze the principal characteristics of the relationship between spot and forward exchange rates which seem to emerge from the experience of the 1970s. Following an analysis of the efficiency of the foreign-exchange market, I discuss the more general

issues underlying the relationships between spot and forward rates and their volatility.

The Efficiency of the Foreign-Exchange Market

One of the central insights of the monetary (or the asset-market) approach to the exchange rate is the notion that the exchange rate, being a relative price of two assets, is determined in a manner similar to the determination of other asset prices and that expectations concerning the future course of events play a central role in affecting current exchange rates.[1]

If the foreign-exchange market is efficient and if the exchange rate is determined in a fashion similar to the determination of other asset prices, we should expect current prices to reflect all currently available information. Expectations concerning future exchange rates should be incorporated and reflected in forward exchange rates. To examine the efficiency of the market, I first regress the logarithm of the current spot exchange rate, $\ln S_t$, on the logarithm of the 1-month forward exchange rate prevailing at the previous month, $\ln F_{t-1}$, as in equation (1):[2]

$$\ln S_t = a + b \ln F_{t-1} + u_t. \tag{1}$$

If the market for foreign exchange is efficient, so that prices reflect all relevant available information, then the residuals in equation (1), u_t, should contain no information and therefore should be serially uncorrelated. Further, if the forward exchange rate is an unbiased forecast of the future spot exchange rate (as should be the case under an assumption of risk neutrality), then the constant term in equation (1) should not differ significantly from zero,[3] and the slope coefficient should not differ significantly from unity. I examine three exchange rates: the dollar/pound, the dollar/franc, and the dollar/DM. Equa-

[1] For collections of articles summarizing this approach, see the *Scandinavian Journal of Economics*, no. 2 (1976) and Frenkel and Johnson (1978).

[2] For an application of the same methodology in analyzing the efficiency properties of the foreign-exchange market during the German hyperinflation of 1921–23 see Frenkel (1976, 1977, 1979). For an application to other exchange rates during the 1920s see Frenkel and Clements (1981); for an application to the 1920s and the 1970s see Krugman (1977); for an interesting analysis using time-series and cross-sectional data see Bilson (in press); for an analysis of market efficiency using novel econometric techniques see Hakkio (1979) and Hansen and Hodrick (1980), and for surveys see Levich (1978, 1979).

[3] More precisely, if (assuming risk neutrality) the forward rate measures the expected value of the future spot rate, then the constant term in the logarithmic equation (1) should be $-.5\sigma_u^2$; see Frenkel 1979. The statement that under risk neutrality the forward rate equals the expected future spot rate neglects the effects of the stochastic elements in prices. As an empirical matter this neglect does not seem to be consequential; see Frenkel and Razin 1980.

tion (1) was estimated using monthly data for the period June 1973–July 1979 (data sources are listed in the Appendix). The beginning of the period was set so as to concentrate on the experience of the current exchange-rate regime (following the initial post–Bretton Woods transition period). The resulting ordinary-least-squares (OLS) estimates are reported in table 1. Also reported in table 1 are additional regressions which will be analyzed shortly.

As may be seen for the dollar/DM exchange rate, the hypotheses that (at the 95 percent confidence level) the constant term does not differ significantly from zero and that the slope coefficient does not differ significantly from unity cannot be rejected. These hypotheses are rejected for the dollar/franc exchange rate and are rejected (marginally) for the dollar/pound exchange rate. The joint hypotheses, however, that the constant is zero and that the slope coefficient is unity cannot be rejected at the 95 percent level for the dollar/pound and the dollar/DM exchange rates and at the 99 percent level for the dollar/franc exchange rate. The test statistics for testing the joint hypotheses are reported in the column headed by F in table 1. These results are relevant for assessing whether the forward rate is an unbiased forecast of the future spot rate. We turn next to the question of efficiency.

It was argued above that in an efficient market, expectations concerning future exchange rates are reflected in forward rates and that spot exchange rates reflect all currently available information. If forward exchange rates prevailing at period $t - 1$ summarize all relevant information available at that period, they should also contain the information that is summarized in data corresponding to period $t - 2$. It thus follows that including additional lagged values of the forward rates in equation (1) should not greatly affect the coefficients of determination and should not yield coefficients that differ significantly from zero. The results reported in table 1 are consistent with this hypothesis; in all cases the coefficients of $\ln F_{t-2}$ do not differ significantly from zero, and the inclusion of the additional lagged variables does not improve the fit. Most important, in all cases the Durbin-Watson statistics are consistent with the hypothesis of the absence of first-order autocorrelated residuals, and an examination of higher-order correlations (up to 12 lags) shows that no correlation of any order is significant.[4]

To examine further the relationship between the various exchange rates, we note that one of the assumptions underlying equation (1)

[4] Since $\ln F_{t-1}$ is highly correlated with $\ln S_{t-1}$ the Durbin-Watson statistic may not be appropriate since equation (1) is very similar to a regression of $\ln S_t$ on its own lagged value. Durbin's h-statistic reveals, however, that the residuals are serially uncorrelated at conventional confidence levels.

TABLE 1

EFFICIENCY OF FOREIGN-EXCHANGE MARKETS
MONTHLY DATA JUNE 1973–JULY 1979
(Standard Errors in Parentheses)

Dependent Variable $\ln S_t$ and Estimation Method	Constant	$\ln F_{t-1}$	$\ln F_{t-2}$	R^2	SE	D-W	F	m
Dollar/pound								
OLS	.033	.956	...	96	.027	1.72	1.86	...
	(.017)	(.024)						
OLS	.031	1.047	−.088	.96	027	1 94	..	...
	(.018)	(.116)	(.113)					
IVAR	030	961	...	.95	.027	1.74	...	2.01
	(.018)	(025)						
Dollar/franc:								
OLS	−.237	843	..	.79	.029	2.23	4.83	...
	(078)	(.051)						
OLS	−.225	706	.146	.79	.029	1.90	. .	...
	(.082)	(.117)	(117)					
IVAR	−.236	.844	..	.78	.030	2.24	...	2.26
	(.080)	(.053)						
Dollar/DM:								
OLS	−.023	.971	. .	.93	.032	2.12	.51	...
	(027)	(.032)						
OLS	−.019	.913	.063	.93	.032	1.96	...	...
	(.028)	(119)	(.122)					
IVAR	−.021	.973	...	.93	.032	2.10	...	.91
	(.027)	(.032)						

NOTE.—SE is the standard error of the equation and R^2 is the coefficient of determination, in the case of instrumental variables estimation, the R^2 was computed as $1 - \text{var}(\hat{u}_t)/\text{var}(\ln S_t)$. The F-statistic tests the joint restriction that the constant equals zero, and the slope equals unity. The test statistic is distributed as $F(2,71)$ Critical values for $F(2,71)$ are 3 13 (95%) and 4.92 (99%). The instrumental variable (IVAR) estimation method is used in order to allow for the possibility of errors in variables arising from using $\ln F_{t-1}$ as a proxy for the expected future spot rate, the instruments are a constant, time time squared, and lagged values of the dependent and the independent variables. The m-statistic which tests for the absence of errors in variables is distributed χ^2 with 2 degrees of freedom The critical value for $\chi^2(2)$ is 5 99 (95%)

was that the forward exchange rate measures the unobservable value of the expected future spot exchange rate. This assumption provided the justification for using equation (1) instead of the explicit specification of the rational expectations hypothesis that is embodied in equation (2):

$$\ln S_t = E_{t-1} \ln S_t + \epsilon_t, \tag{2}$$

where $E_{t-1} \ln S_t$ denotes the expected (logarithm of the) spot exchange rate for period t based on the information available at period $t - 1$. If, however, the forward exchange rate at $t - 1$ is a "noisy" proxy for the expected future value of the spot rate (i.e., it measures it with a random error), then we would obtain

$$a + b \ln F_{t-1} = E_{t-1} \ln S_t + v_{t-1}; E(v_t) = 0, \tag{3}$$

and substituting equation (3) into equation (2) yields:

$$\ln S_t = a + b \ln F_{t-1} + (\epsilon_t - v_{t-1}). \tag{4}$$

In this case the error term in equation (1) would be $u_t = \epsilon_t - v_{t-1}$; the assumption that the covariance between $\ln F_{t-1}$ and u_t is zero would entail a specification error, and the application of the OLS procedure would yield biased estimates due to the classical errors in variables bias.

In order to examine the possibility that the OLS estimates might be subject to the errors in variables bias, one needs to test the hypothesis that cov $(u_t, \ln F_{t-1}) = 0$. This test follows the specification test outlined by Hausman (1978).[5] To perform the test, equation (1) was estimated by applying the OLS procedure as well as by using an instrumental variables (IVAR) estimation method. Under the null hypothesis of no misspecification the OLS coefficients vector $\hat{b}_0$ is an efficient and unbiased estimate of the true coefficient vector. Under the alternative hypothesis of misspecification, the vector $\hat{b}_0$ is biased and an unbiased coefficient vector $\hat{b}_1$ can be obtained by applying an instrumental variables estimation procedure. The test statistic relevant for testing the null hypothesis can be written as

$$m = (\hat{b}_1 - \hat{b}_0)'(\text{var } \hat{b}_1 - \text{var } \hat{b}_0)^{-1}(\hat{b}_1 - \hat{b}_0), \tag{5}$$

where var $(\hat{b}_1)$ and var $(\hat{b}_0)$ denote the variance-covariance matrices of $\hat{b}_1$ and $\hat{b}_0$, respectively. Under the null hypothesis, m is distributed (in large samples) as χ^2 with 2 degrees of freedom. Table 1 reports the results of estimating equation (1) by applying the instrumental variables estimation method. As may be seen for all exchange rates the

[5] This test was recently applied by Obstfeld (1978) to the analysis of the foreign-exchange market during the 1970s and by Frenkel (1980*a*, 1980*b*) to the analysis of the foreign-exchange markets during the 1920s.

two vectors of coefficients $\hat{b}_1$ and $\hat{b}_0$ are very close to each other. For example, for the dollar/pound exchange rate, the constants are .033 and .030, and the slopes are .956 and .961—consequently, the resulting m-statistic is 2.01, which is well below 5.99—the critical value of $\chi^2(2)$ at the 95 percent confidence level. The m-statistics corresponding to the other exchange rates are also below this critical value. It is concluded, therefore, that the use of the forward exchange rate as a proxy for expectations does not introduce a significant errors in variables bias and, thus, the use of the OLS estimation procedure seems appropriate.[6]

The principal conclusion that may be drawn from the previous discussion is that the behavior of the foreign-exchange market during the 1970s has been broadly consistent with the general implications of the efficient market hypothesis.

Exchange-Rate Movement: Volatility and Predictability

In this section I analyze the volatility of exchange rates and the extent to which this volatility is predictable. As is well known, during the same period exchange rates have been very volatile. The standard errors of the monthly percentage changes of the three exchange rates have been about 3 percent per month. Further, the standard errors of the regressions in table 1 indicate that the forecasts of future spot exchange rates based on the forward rates are very imprecise: The standard errors of the equations are about 3 percent per month.

[6] The efficiency of the foreign-exchange market can also be analyzed from a different angle as in Frenkel (1980*b*). Consider the equation

$$x_t = \alpha_0 + \alpha_1 t + \sum_{i=1}^{n} \beta_i x_{t-i} + \gamma \pi_{t-1} + w_t,$$

where x_t denotes the percentage change of the spot exchange rate ($\ln S_t - \ln S_{t-1}$); π_{t-1} denotes the forward premium on foreign exchange ($\ln F_{t-1} - \ln S_{t-1}$); t denotes time; n denotes the number of lags; and w denotes an error term. If π_{t-1} summarizes all available information concerning the future evolution of the exchange rate, then given the value of the forward premium π_{t-1}, the past history of the percentage change of the exchange rate should not "help" the prediction (i.e., the past history should not be viewed as Granger-causing future changes), and the joint hypotheses that α_0, α_1, and β_i are zero and that γ is unity should not be rejected. The results of applying these tests to the three exchange rates for various numbers of lags as well as to the pooled data base of the three exchange rates show that the null hypothesis cannot be rejected at the 95 percent confidence level, since the values of the various F-statistics fall well below the corresponding critical values. It is noteworthy, however, that the power of this test is low and that the joint hypothesis that α_0, α_1, β_i, and γ are zero could also not be rejected. The difficulties in "explaining" the percentage change of the exchange rate in terms of past values of various variables reflect the fact that like the prices of other assets which are traded in organized markets, changes in exchange rates are dominated by "news" which by definition could not have been incorporated in past changes or in the lagged forward premium. For a further elaboration see the following section.

These characteristics of price changes (volatility and unpredictability) are typical of auction and of organized asset markets. In such markets current prices reflect expectations concerning the future course of events, and new information which induces changes in expectations is immediately reflected in corresponding changes in prices, thus precluding unexploited profit opportunities from arbitrage. The strong dependence of current prices on expectations about the future is unique to the determination of durable asset prices; it is not a characteristic of price determination of nondurable commodities. The strong dependence of asset prices on expectations also implies that periods which are dominated by uncertainties, new information, rumors, announcements, and "news," which induce frequent changes in expectations, are likely to be periods in which changes in expectations are the prime cause of fluctuations in asset prices. Further, since the information which alters expectations must be new, the resulting fluctuations in price cannot be predicted by lagged forward exchange rates which are based on past information.[7] Therefore, during such periods, one should expect exchange rates (and other asset prices) to exhibit large fluctuations. When the prime cause of fluctuations is new information, one may expect that lagged forward exchange rates (which are based on past information) are imprecise (even though possibly the best unbiased) forecasts of future spot rates.

To gain further insights into the implications of this perspective on the relationship between predicted and realized changes in exchange rates, figures 1–3 present plots of predicted and realized percentage changes in exchange rates for the three pairs of currencies where the predicted change is measured by the lagged forward premium. Also presented in these figures are the differentials in national inflation rates which are discussed in Section IV. The key fact which emerges from these figures is that predicted changes in exchange rates account for a very small fraction of actual changes.[8] This phenomenon is also reflected in the comparison between the variances of actual and predicted changes in exchange rates: In all cases the variances of monthly percentage changes in exchange rates exceed the variances of monthly forward premia by a factor that is larger than 20.[9] This fact suggests that the bulk of exchange-rate changes seem to be due to

[7] The analysis of the role of "news" in determining current exchange rates and in explaining forecast errors from the forward rate has been made forcefully by Mussa (1976*a*, 1976*b*, 1977, 1979*a*) and Dornbusch (1978). The large degree of volatility is also analyzed by McKinnon (1976), who attributes it to insufficient speculation.

[8] These and the following empirical regularities are analyzed in detail in Mussa (1979*a*). See also Frenkel and Mussa 1980.

[9] For an analysis of the relationship between the variances of series of predictions and series of realizations see Shiller (1979) and Singleton (1980).

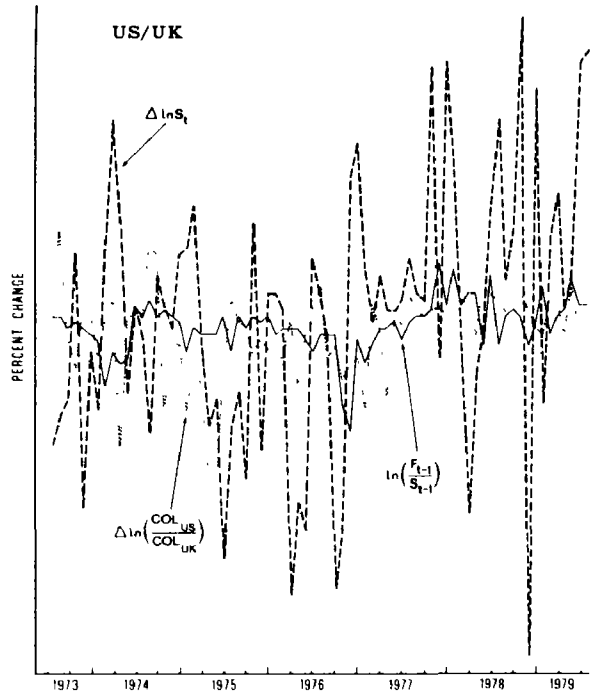


FIG. 1.—Monthly percentage changes of the U.S./U.K. consumer price indices [$\Delta(\ln COL_{US}/COL_{UK})$]; of the \$/£ exchange rate ($\Delta \ln S_t$); and the monthly forward premium [$\ln(F_{t-1}/S_{t-1})$], July 1973–July 1979

"new information" which, by definition, could not have been anticipated and reflected in the forward premium or discount which prevailed in the previous period.

This view of the foreign-exchange market can be exposited in terms of the following simple model.[10] Let the logarithm of the spot exchange rate on day t be determined by

$$\ln S_t = z_t + bE_t(\ln S_{t+1} - \ln S_t), \tag{6}$$

where $E_t(\ln S_{t+1} - \ln S_t)$ denotes the expected percentage change in the exchange rate between t and $t + 1$, based on the information available at t, and where z_t represents the ordinary factors of supply and demand that affect the exchange rate on day t. These factors may include domestic and foreign money supplies, incomes, levels of output, etc. Equation (6) represents a sufficiently general relationship which may be viewed as a "reduced form" that can be derived from a variety of models of exchange-rate determination. These models may

[10] The following paragraph draws on Frenkel and Mussa (1980)

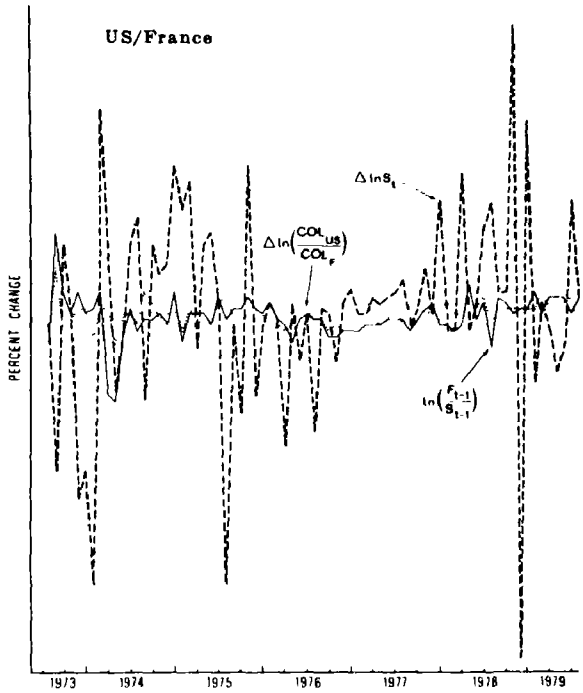


FIG. 2.—Monthly percentage changes of the U.S./French consumer price indices [$\Delta(\ln COL_{US}/COL_F)$]; of the \$/Fr exchange rate ($\Delta \ln S_t$); and the monthly forward premium [$\ln(F_{t-1}/S_{t-1})$], July 1973–July 1979.

differ in their emphasis of the determinants of $z(t)$, but they all are likely to share a similar reduced form.[11] Assuming that expectations are rational in that equation (6) applies to expectations of future exchange rates, it follows, by forward iteration, that

$$E_t \ln S_{t+j} = \frac{1}{1+b} \sum_{k=0}^{\infty} \left(\frac{b}{1+b}\right)^k E_t z_{t+j+k}. \tag{7}$$

Thus, the current exchange rate ($j = 0$) and current expectations of future exchange rates ($j > 0$) are linked, because both depend on expectations concerning the future z's. The strength of the link depends on the magnitude of b which characterizes the dependence of the current exchange rate on the expected percentage change thereof.[12] The presumption is that, due to profit opportunities from

[11] See, e.g., the comprehensive econometric model of Fair (1979).

[12] A result of this general form is derived in Mussa (1976*a*). The unique role of expectations is also emphasized by Black (1973), Dornbusch (1976*c*, 1978), Kouri (1976), and Bilson (1978). In general, the value of b may be viewed as the relevant

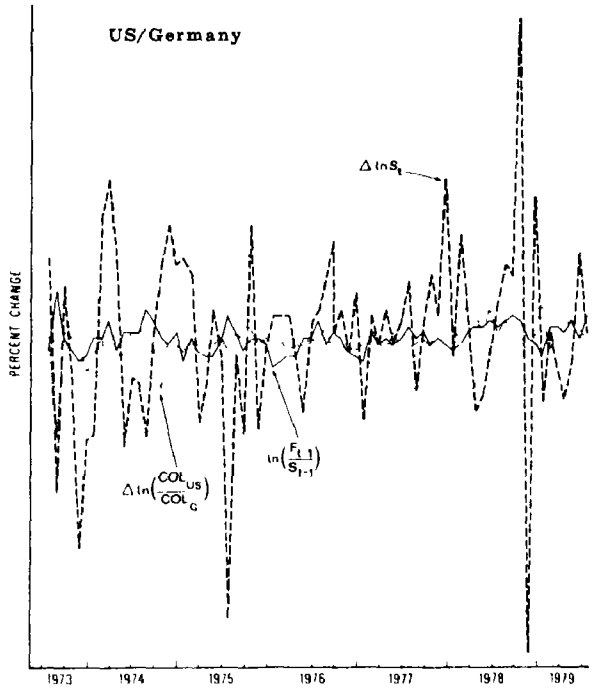


FIG. 3.—Monthly percentage changes of the U.S./German consumer price indices [$\Delta(\ln COL_{US}/COL_G)$]; of the \$/DM exchange rate ($\Delta \ln S_t$); and the monthly forward premium [$\ln(F_{t-1}/S_{t-1})$], July 1973–July 1979

arbitrage, this link is strong at least for the exchange rates expected in the near future. Hence, the current exchange rate, $\ln S_t = E_t \ln S_t$, should be closely linked to the current expectation of the next period's exchange rate, $E_t \ln S_{t+1}$, which in turn should be closely linked to the exchange rate expected for the following period, $E_t \ln S_{t+2}$, and so on.

In order to examine this hypothesis I present in figures 4–6 plots of the spot and the contemporaneous forward exchange rates for the three pairs of currencies. Also presented are the ratios of national price levels which are discussed in Section IV. If the dominant factor underlying changes in rates is new information which alters views about current and expected future exchange rates by approximately the same amount, then one should expect a high correlation between movements of spot and forward rates. This fact is clearly demon-

parameter for determining whether a specific commodity (whose pricing rule is described in terms of eqq. like [6]–[7]) may be viewed as an asset. The higher the value of b is for a given commodity, the larger is its asset attribute.

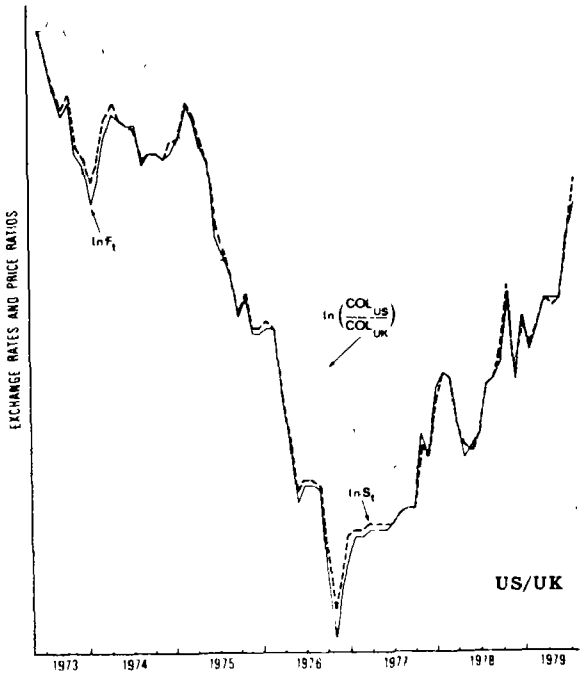


FIG 4.—Monthly observations of the \$/£ spot ($\ln S_t$) and forward ($\ln F_t$) exchange rates and the ratio of the U.S./U K. cost of living indices [$\ln(COL_{US}/COL_{UK})$] (scaled to equal the spot exchange rate at the initial month): June 1973–July 1979.

strated by figures 4–6 where it is seen that spot and forward exchange rates tend to move together and by approximately the same amount (the vertical difference between the two rates corresponds to the percentage forward premium or discount on foreign exchange). The correlations between the spot and the forward exchange rates for the three pairs of currencies exceed .99, and the correlations between the corresponding percentage changes of the spot and forward rates exceed .96. The high correlation between movements in spot and forward rates is expected, since the two rates respond at the same time to the same flow of new information (which is presumed to affect the rate for more than one period). In general, the details of the relationship between spot and forward exchange rates depend on the time-series properties of the z's in equation (7) and in particular on whether the new information is viewed as permanent or transitory.[13]

[13] New information might be "permanent" when the relevant horizon is 1 month, while it might be transitory when the relevant horizon is a year. In that case, the correlation between the spot exchange rate and the contemporaneous 1-month for-

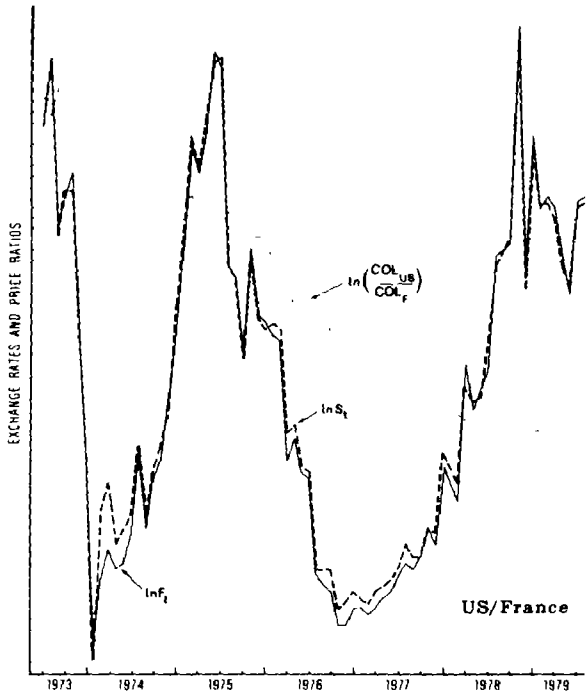


FIG. 5.—Monthly observations of the \$/Fr spot ($\ln S_t$) and forward ($\ln F_t$) exchange rates and the ratio of the U.S./French cost-of-living indices [$\ln(COL_{US}/COL_F)$] (scaled to equal the spot exchange rate at the initial month): June 1973–July 1979.

The comovement of spot and forward rates is evidence of the close link between current and expected future exchange rates which is illustrated by equation (7). This characteristic is typical of the foreign-exchange market and is also shared by prices of many assets and commodities traded in organized markets. The recent pattern of gold prices provides a useful example of this general principle. Table 2 reports the spot and the futures prices of gold as recorded recently in the International Money Market (IMM) at the Chicago Mercantile Exchange on 5 consecutive days. The two key facts which are illustrated by this table are (i) the extent of day-to-day volatility in gold prices and (ii) the general uniformity by which these changes are reflected in the price of gold for immediate delivery as well as in the prices for the eight future delivery dates.

ward rate is likely to be high, while the correlation between the spot rate and the 12-month forward rate is likely to be low. The perceived permanence of the new information can be inferred from the correlation between the spot and the various maturities of the forward rate. As expected, it is generally found that this correlation diminishes with the maturity of the forward contract.

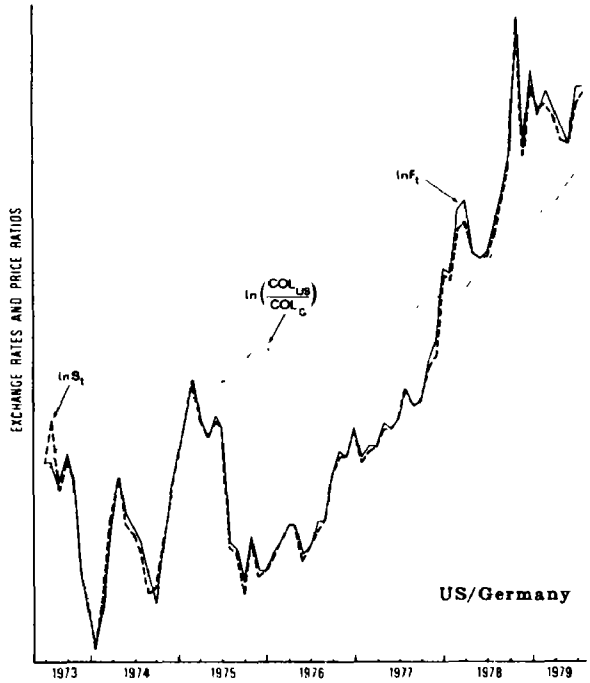


FIG. 6.—Monthly observations of the \$/DM spot ($\ln S_t$) and forward ($\ln F_t$) exchange rates and the ratio of the U.S./German cost-of-living indices [$\ln(COL_{US}/COL_G)$] (scaled to equal the spot exchange rate at the initial month): June 1973–July 1979.

Another feature which is revealed by figures 4–6 is that the contemporaneous spot and forward exchange rates are approximately equal, thus indicating that the market's best forecast of the future spot rate is (approximately) the current spot rate. This phenomenon reflects the fact that, as an empirical matter, exchange rates have followed (approximately) a random-walk process. For such a process, current prices are, indeed, the best forecasts of future prices (and to the extent that the exchange rate had some drift, the statement above should be interpreted in reference to that drift). It is relevant to note, however, that while the random-walk phenomenon seems to correspond to the actual paths of exchange rates, it does not reflect a theoretical necessity.

The final characteristic of the foreign-exchange market is described by figures 7–9, which plot for the three pairs of currencies the spot exchange rate and the forward premium on forward exchange. Since the spot rate and the forward premium are expressed in terms of different units (where the latter is expressed as percentages per

TABLE 2

FUTURES PRICE OF GOLD ON CONSECUTIVE DAYS
DAILY DATA: MAY 1, 1980–MAY 7, 1980

DELIVERY DATE	PRICE PER OUNCE AND CHANGE FROM PREVIOUS DAY									
	May 1	Change	May 2	Change	May 5	Change	May 6	Change	May 7	Change
1980:										
May	516.0	14.4	511.0	−5.0	519.0	8.0	506.5	−12.5	512.0	5.5
June	517.0	10.5	575.5	−1.5	523.7	8.2	510.3	−13.4	516.0	5.7
September	536.0	10.5	533.5	−2.5	540.7	7.2	527.0	−13.7	531.5	4.5
December	554.0	9.5	551.0	−3.0	558.0	7.0	541.5	−16.5	547.0	5.5
1981:										
March	571.5	9.5	568.0	−3.5	574.7	6.7	557.0	−17.7	561.5	4.5
June	588.5	10.5	585.0	−3.5	591.0	6.0	571.5	−19.5	576.0	4.5
September	605.3	10.3	602.0	−3.3	607.0	5.0	585.5	−21.5	590.0	4.5
December	622.0	11.0	619.0	−3.0	623.0	4.0	599.5	−23.5	604.0	4.5
1982:										
March	638.5	11.5	636.0	−2.5	693.0	3.0	613.5	−25.5	618.0	4.5

NOTE.—These prices are settlement prices at the International Money Market (IMM) at the Chicago Mercantile Exchange as reported in the *Wall Street Journal* (May 2–May 8, 1980).

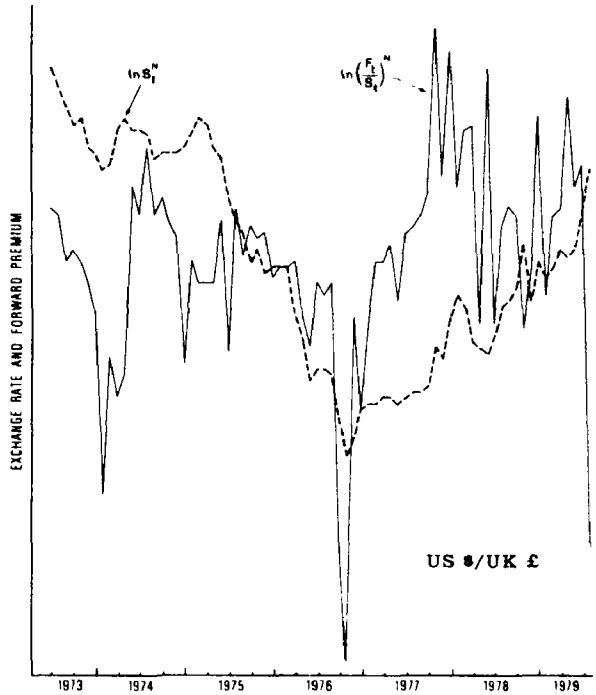


FIG. 7.—Monthly observations of the normalized $/£ spot exchange rate ($\ln S_t^N$) and the normalized forward premium [$\ln(F_t/S_t)^N$]. Both series are normalized by subtracting from each series its mean and by dividing by the corresponding standard error. June 1973–July 1979.

month), the two series were normalized for the purpose of the plots by subtracting from each series its mean and by dividing by the corresponding standard error.[14] The fact which emerges from these figures is that generally (though not always) there is a positive correlation between the expected depreciation of the currency (as measured by the forward premium on foreign exchange) and the spot exchange rate. This positive correlation may be rationalized by noting that currencies which are expected to depreciate are traded at a discount in the forward market and, on average, these currencies also command a lower foreign-exchange value in the spot market.[15] This

[14] The normalized values of the spot rate, $\ln S_t^N$, and of the forward premium, $\ln(F_t/S_t)^N$, which are plotted in figs. 7–9, are defined as $\ln S_t^N = (\ln S_t - \overline{\ln S})/\sigma_{\ln S}$, and $\ln(F_t/S_t)^N = [\ln(F_t/S_t) - \overline{\ln(F/S)}]/\sigma_{\ln(F/S)}$, where a bar over a variable indicates its sample mean, and where σ denotes the sample standard deviation of the subscripted variable.

[15] It is noteworthy that since the forward premium (like the rate of interest) and the exchange rate are dimensionally incommensurate, their association raises questions

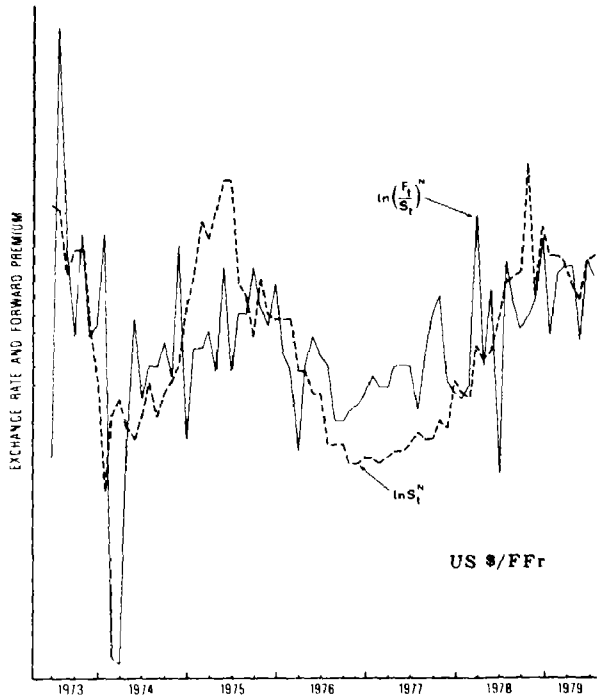


FIG. 8.—Monthly observations of the normalized $/Fr spot exchange rate ($\ln S_t^N$) and the normalized forward premium [$\ln(F_t/S_t)^N$]. Both series are normalized by subtracting from each series its mean and by dividing by the corresponding standard error: June 1973–July 1979

relationship is embodied in the specification of equation (6) and is interpreted further in the next section.

III. Exchange Rates, Interest Rates, and Innovations

This section contains an analysis of the relationship between exchange rates and interest rates from the perspective of the monetary (or the asset-market) approach to the exchange rate. Following a discussion of the broad facts, the analysis proceeds with an empirical examination of the role of "news."

Exchange Rates and Interest Rates: The Broad Facts

To set the stage for this section, it is useful to recall the analysis which predicts a negative association between the rate of interest and the

that are familiar from the discussions of the Gibson paradox. In a separate paper I intend to examine the relationship between exchange rates and the forward premium (or the interest differential) in light of the various explanations of the Gibson paradox.

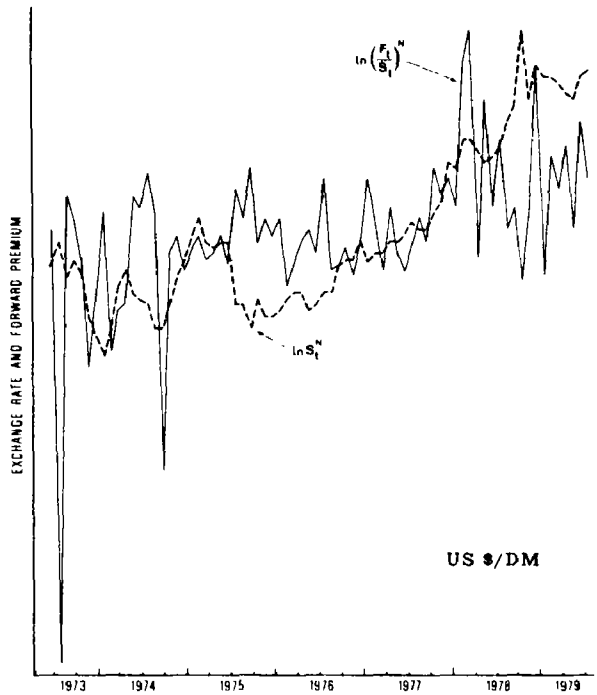


FIG. 9.—Monthly observations of the normalized $/DM spot ($\ln S_t^N$) and the normalized forward premium [$\ln(F_t/S_t)^N$]. Both series are normalized by subtracting from each series its mean and by dividing by the corresponding standard error: June 1973–July 1979.

exchange rate. According to that analysis, a higher rate of interest attracts foreign capital which induces a surplus in the capital account of the balance of payments and thereby induces an appreciation of the domestic currency (i.e., a lower spot exchange rate). Another variant of this approach states that the higher rate of interest lowers spending and thus induces a surplus in the current account of the balance of payments which results in a lower spot exchange rate. A third variant claims that the higher rate of interest implies (via the interest parity theory) a higher forward premium on foreign exchange; and to the extent that at a given point in time the forward exchange rate which represents the expected future spot rate is predetermined by past history—as would be the case under the adaptive expectations hypothesis (which is clearly rejected by the evidence on the comovements of spot and forward rates)—the required rise in the forward premium will be brought about by a lower spot rate (i.e., by an appreciation of the domestic currency). Whatever the route, by ignoring the distinction between nominal and real rates of interest, this approach predicts a *negative* relationship between the rate of

interest and the spot exchange rate (or, alternatively, a positive relationship between the rate of interest and the foreign-exchange value of the domestic currency).

While such a prediction might be appropriate for noninflationary environments, it might be entirely inappropriate for inflationary environments (like the one prevailing in the United States in recent years). In such periods variations in rates of interest are most likely to be dominated by variations in inflationary expectations rather than by liquidity effects associated with changes in the ratio of money to bonds. In such an environment, the rate of interest is expected to be *positively* correlated with the exchange rate. The broad facts are consistent with this hypothesis. Over the recent period, the rise in the rate of interest in the United States (relative to the foreign rate of interest) has been associated with a rise in the spot exchange rate (i.e., with a depreciation of the dollar) rather than with a fall in the spot rate. Figure 10 illustrates the point by plotting the foreign-exchange value of the U.S. dollar against the interest rate differential. As is evident, the higher (relative) rate of interest in the United States has been associated with a higher exchange rate (i.e., with a lower foreign-exchange value of the dollar).

The positive association between the rate of interest and the exchange rate in the context of the U.S. dollar and the inflationary environment can be accounted for by the monetary (or the asset-market) approach to the exchange rate which puts a special emphasis on the influence of the expectations on the current values of exchange rates.[16] For example, according to the monetary approach, a rise in the domestic (relative) rate of interest which is primarily dominated by a rise in the expected (relative) rate of inflation induces a decline in the demand for real cash balances; for a given path of the nominal money supply, asset-market equilibrium requires a price level which is higher than the price which would have prevailed otherwise. When the domestic price level is linked to the foreign price through some form of purchasing power parity, and when the path of the foreign price is assumed to be given, the higher domestic price can only be achieved through a rise in the spot exchange rate (i.e., through a depreciation of the currency).[17]

[16] For theoretical developments and applications of the approach see, e.g., Dornbusch (1976*a*, 1976*b*), Frenkel (1976), Kouri (1976), Mussa (1976*a*), Bilson (1978), Frenkel and Johnson (1978), Hodrick (1978), Frankel (1979), Clements and Frenkel (1980), and Frenkel and Clements (1981).

[17] It should be emphasized that this explanation of the positive association between the rate of interest and the exchange rate does not rely on a rigid form of the purchasing power parity theory. It only requires that domestic and foreign price levels, when expressed in terms of the same currency, are positively correlated. The evidence from the 1970s is consistent with this requirement; see Frenkel 1981.

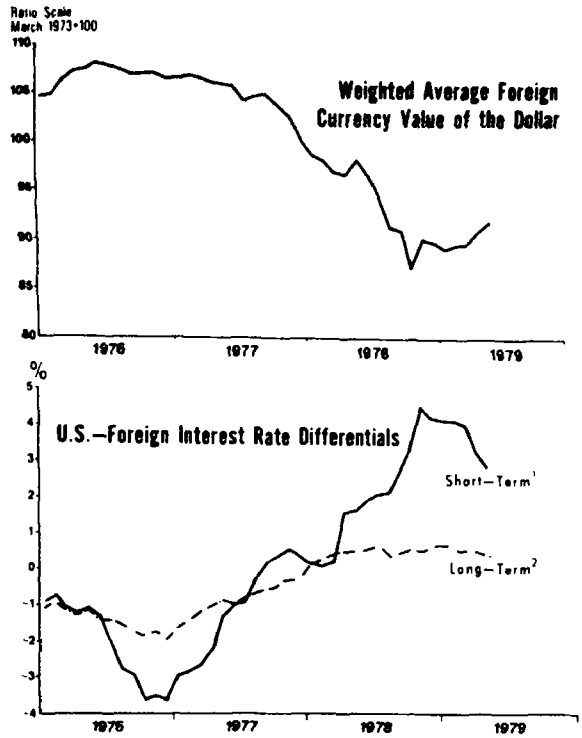


FIG. 10.—Foreign-exchange value of the U.S. dollar and interest rate differentials
Source: Mudd 1979.

This explanation of the positive association between interest rates and exchange rates has an intuitive appeal in that it implies that, in an inflationary environment, a relatively rapid rise in prices is associated with high *nominal* rates of interest as well as with a depreciation of the currency in terms of foreign exchange. This relationship is embodied in equation (6), which states that an expected depreciation of the currency (which in our case is associated with inflationary expectations and high nominal rates of interest) results in an immediate depreciation.[18]

The foregoing analysis also provides the explanation for the obser-

[18] The traditional prediction of a negative relationship between interest rates and the exchange rate could, in principle, be rationalized under the assumption that it concentrates on the short-run liquidity effects of monetary changes. It should be emphasized, however, that in an inflationary environment, like the one prevailing in the United States, the applicability of this rationalization is very limited. The short-run liquidity effect is emphasized in Dornbusch (1976*b*). The role of inflationary expectations in dominating exchange-rate developments is emphasized in Frenkel (1976). Edwards (1979) and Frankel (1979) attempt to integrate these two factors.

vation (which was noted in Sec. II) that, generally, there is a positive correlation between the forward premium on foreign exchange and the level of the spot rate. Since during inflationary periods the spot rate is expected to be positively correlated with the interest rate differential, and since according to the interest parity theory that differential must equal the forward premium on foreign exchange, it follows that the forward premium is also expected to be positively correlated with the level of the spot rate.[19]

Exchange Rates and "News"

One of the central implications of the rational expectations hypothesis is that unanticipated events, "news," play a predominant role in affecting real variables and asset yields. This implication is embodied in many expositions of modern macroeconomics, and its empirical content has been the subject of numerous recent studies.[20] In the context of exchange-rate determination the discussion in Section II and in particular the contributions by Mussa (1977, 1979*a*) and Dornbusch (1978) emphasized that the predominant cause of exchange-rate movements is "news" which could not have been anticipated.[21] Expressing the spot exchange rate at period t as the sum of factors which were anticipated from the past, as well as factors which represent "news," Dornbusch (1978) decomposes the effects of "news" into those which alter the expected future spot rate between the last period and the present and those which lead to a reassessment of the 1-period interest rate differential starting at the present, that is, "news" about the term structure. Both of these white-noise, serially uncorrelated components play a role in determining the spot exchange rate in Dornbusch's analysis.

The evidence presented in Section II suggests that the forward rate summarizes the information that is available to the market when the forward rate is being set, and in equation (3) it was assumed that the expected exchange rate can be written as $a + b \ln F_{t-1}$ plus a serially uncorrelated error. We may, therefore, express the spot rate at period t as a function of factors which have been known in advance and are summarized by the lagged forward rate, as well as a function of the "news":

$$\ln S_t = a + b \ln F_{t-1} + \text{"news"} + w_t. \tag{8}$$

In what follows this notion is applied to an empirical analysis of the

[19] For evidence on the robustness of the interest parity relationship see Frenkel and Levich (1977).

[20] See, e.g., Barro 1977 and Fischer 1980.

[21] See also Bilson 1978; Frenkel and Mussa 1980; Isard 1980; and Longworth 1980.

role of "news" as a determinant of the exchange rate. The key difficulty lies in identifying the variable which measures the "news." Since, quite frequently, it is difficult to observe and quantify the "news," it is convenient to examine the relationship between the exchange rate and a variable whose time series is likely to manifest the "news" promptly. Assuming that asset markets clear fast and that the "news" is immediately reflected in (unexpected) changes in the rates of interest, equation (8) may be written as

$$\ln S_t = \underbrace{a + b \ln F_{t-1}}_{\text{expected exchange rate}} + \underbrace{\alpha[(i - i^*)_t - E_{t-1}(i - i^*)_t]}_{\text{"news"}} + w_t, \qquad (9)$$

where the bracketed term denotes the innovation in the (1-month) interest differential and where $E_{t-1}(i - i^*)_t$ denotes the interest differential which was expected to prevail in period t based on the information available at $t - 1$. The expected interest rate differential was computed from a regression of the interest differential on a constant, on two lagged values of the differential, and on the lagged forward exchange rate $\ln F_{t-1}$.

As was argued above, the association between exchange rates and interest rates is likely to be positive during periods in which most of the variations in nominal rates of interest are dominated by variations in inflationary expectations—a characteristic which seems to fit the inflationary environment of the 1970s. Under such circumstances, when the unexpected interest differential reflects "news" concerning inflationary expectations, the coefficient α in equation (9) is likely to be positive.[22] Table 3 reports the two-stage least-squares estimates of equation (9) for the three exchange rates over the period June 1973–July 1979.[23] As may be seen in all cases, the coefficients of the unexpected interest differential are positive and, in the case of the dollar/pound exchange rate, the coefficient is statistically significant. In

[22] In general, of course, the sign of the coefficient α depends on the source of the variation in the interest rate.

[23] In all cases, the lagged forward exchange rate was included as an instrument in order to obtain consistent estimates; see Nelson 1975. Adding lagged values of the percentage changes of the domestic and the foreign money supplies as determinants of the expected interest differential and adding the current values of the percentage change of the money supplies as instruments for the unexpected interest differential did not affect the results in any material way. In order to obtain consistent estimates, the assumption made in the regressions reported in tables 3–6 is that, in forming expectations concerning the interest differential, individuals look only at the lagged forward premium and at past values of the differential. An alternative way to compute the expected differential would use data on the term structure of interest rates. Since data on the differential of 2-month rates are not readily available, this computation would require interpolations.

order to verify the importance of using the series of innovations in the interest differential, table 3 also reports estimates of regressions which replace the innovations by the actual series of the interest differential. In all cases the coefficients of the actual interest differential do not differ significantly from zero. Table 4 describes analogous two-stage least-squares estimates of regressions which include both the actual interest rate differential and the innovation in the differential.[24] Again, in all cases the coefficients on the actual differential do not differ significantly from zero, while the coefficients on the innovations are all positive and are significant for the dollar/pound and the dollar/franc exchange rates.

One possible interpretation of the positive coefficient on the (unexpected) interest differential may be given in terms of the prediction of the monetary approach to the exchange rate, in which case the estimate of the coefficient α in equation (9) might be interpreted as an estimate of a structural parameter. Under an alternative interpretation the innovations in the interest differential belong in equation (9) only as far as they manifest "news" which is relevant for exchange-rate determination. If, for example, the dominant element of "news" was variation in inflationary expectations, then one could also use the innovations in other time series as long as they reflect the relevant "news." To examine this possibility the same regressions as in tables 3–4 were estimated using the 12-month interest differential, and the results are reported in tables 5–6. As before, in all cases the coefficients on the actual differential do not differ significantly from zero, while in all cases the coefficients on the innovations in the differential are positive and significant for the dollar/pound and the dollar/franc exchange rates.[25]

On the whole the record shows that during the 1970s exchange

[24] The difficulties in obtaining instruments for data that are innovations are obvious since, by virtue of being "news," it is unlikely that variables which characterize the history can serve as good instruments. The difficulties are acute in cases where a variable and its expected value appear in the same regression, in which case consistent estimates require the use of an instrument that is contemporaneous with the innovation and is exogenous (see McCallum 1979). Thus, in addition to a constant and the lagged forward exchange rate, Durbin's rank variable was used as an instrument for estimating the innovations in the interest rate differential.

[25] To allow for the possibility that the exchange-rate equation includes both the short- and the long-term interest rates where, as suggested by Frankel (1979), the former captures liquidity effects and the latter captures expectations effects, eq. (9) was also estimated using the innovations in both the 1-month and the 12-month interest differential on the right-hand side. Since the two sets of innovations are highly collinear, none of the coefficients differed significantly from zero. In this context it is also noteworthy that the 12-month interest rate contains elements of the 1-month rates, due to the characteristics of the term structure of interest rates. As a result, coefficient estimates from regressions which use both rates, like those in Frankel (1979), must be interpreted with great care.

TABLE 3

ONE-MONTH INTEREST RATE DIFFERENTIALS AND EXCHANGE RATES; INSTRUMENTAL VARIABLES
MONTHLY DATA. JUNE 1973–JUNE 1979
(Standard Errors in Parentheses)

Dependent Variable ln S_t	Constant	ln F_{t-1}	$(i - i^*)_t$	$[(i - i^*)_t - E_{t-1}(i - i^*)_t]$	SE	R^2	D-W
Dollar/pound	.021	.965	− 152	..	.028	.95	1.69
	(020)	(026)	(.118)				
	.031	.959	...	.432	.026	.96	1.78
	(.017)	(.024)		(.181)			
Dollar/franc	−.024	992	−.462	...	.034	.71	2.21
	(.181)	(124)	(.324)				
	− 246	.837	...	.245	.029	.80	2.17
	(077)	(.051)		(.167)			
Dollar/DM	.004	997	−.180	...	.033	.93	2.12
	(.064)	(064)	(.394)				
	−.022	.972	..	.413	.031	.93	2.05
	(.026)	(.031)		(.347)			

NOTE —Interest rates are the 1-month (annualized) Euromarket rates The expected interest rate differential $E_{t-1}(i - i^*)_t$ was computed from a regression of the interest differential on a constant, two lagged values of the differential, and the logarithm of the lagged forward exchange rate A two-stage least-squares estimation method was used The instruments for the interest differential were a constant, two lagged values of the differential, and the logarithm of the lagged forward exchange rate, the instruments for the unexpected differential were a constant, Durbin's rank variable, and the logarithm of the lagged forward exchange rate. The actual interest rate differential is denoted by $(i - i^*)_t$, where i denotes the rate of interest on securities denominated in U.S. dollars, and i^* denotes the rate of interest on securities denominated in foreign currency The unexpected interest rate differential is denoted by $[(i - i^*)_t - E_{t-1}(i - i^*)_t]$. SE is the standard error of the equation. A quasi-R^2 was computed as $1 - \text{var}(\hat{u}_t)/\text{var}(\ln S_t)$

TABLE 4

ONE-MONTH INTEREST RATE DIFFERENTIALS AND FORECAST ERRORS OF EXCHANGE RATES; INSTRUMENTAL VARIABLES
MONTHLY DATA: JUNE 1973–JULY 1979
(Standard Errors in Parentheses)

Dependent Variable $\ln S_t$	Constant	$\ln F_{t-1}$	$(i - i^*)_t$	$[(i - i^*)_t - E_{t-1}(i - i^*)_t]$	SE	R^2	D-W
Dollar/pound	.018	.969	−.156	.562	.023	.97	1.81
	(.017)	(.022)	(.100)	(.191)			
Dollar/franc	−.136	.915	−.312	.547	.031	.75	2.19
	(.112)	(.076)	(.209)	(.282)			
Dollar/DM	.002	.996	−.173	.599	.032	.93	2.07
	(.044)	(.045)	(.286)	(.457)			

NOTE.—Interest rates are the 1-month (annualized) Euromarket rates. The expected interest rate differential $E_{t-1}(i - i^*)_t$ was computed from a regression of the interest differential on a constant, two lagged values of the differential, and the logarithm of the lagged forward exchange rate. A two-stage least-squares estimation method was used. The instruments were a constant, two lagged values of the interest differential, Durbin's rank variable of the unexpected differential, and the logarithm of the lagged forward exchange rate. The actual interest rate differential is denoted by $(i - i^*)_t$, where i denotes the rate of interest on securities denominated in U.S. dollars, and i^* denotes the rate of interest on securities denominated in foreign currency. The unexpected interest rate differential is denoted by $[(i - i^*)_t - E_{t-1}(i - i^*)_t]$. SE is the standard error of the equation. A quasi-R^2 was computed as $1 - \text{var}(\hat{u}_t)/\text{var}(\ln S_t)$.

TABLE 5

TWELVE-MONTH INTEREST RATE DIFFERENTIALS AND EXCHANGE RATES; INSTRUMENTAL VARIABLES
MONTHLY DATA: JUNE 1973–JULY 1979
(Standard Errors in Parentheses)

Dependent Variable $\ln S_t$	Constant	$\ln F_{t-1}$	$(i - i^*)_t$	$[(i - i^*)_t - E_{t-1}(i - i^*)_t]$	SE	R^2	D-W
Dollar/pound	.027	.960	−.083	...	.027	.95	1.71
	(.019)	(.025)	(.147)				
	.031	.959	...	.887	.025	.96	1.81
	(.017)	(.024)	...	(.291)			
Dollar/franc	−.152	.904	−.230	...	.030	.77	2.32
	(.155)	(.107)	(.320)				
	−.246	.837	...	.729	.029	.80	2.10
	(.077)	(.051)		(.342)			
Dollar/DM	−.012	.982	−.074	...	.032	.93	2.11
	(.063)	(.063)	(.395)				
	−.022	.972	...	.979	.031	.94	2.07
	(.026)	(.031)		(.663)			

NOTE.—Interest rates are the 12-month Euromarket rates. The expected interest rate differential $E_{t-1}(i - i^*)_t$ was computed from a regression of the interest differential on a constant, two lagged values of the differential, and the logarithm of the lagged forward exchange rate. A two-stage least-squares estimation method was used. The instruments for the interest differential were a constant, two lagged values of the differential, and the logarithm of the lagged forward exchange rate; the instruments for the unexpected differential were a constant, Durbin's rank variable, and the logarithm of the lagged forward exchange rate. The actual interest rate differential is denoted by $(i - i^*)_t$, where i denotes the rate of interest on securities denominated in U.S. dollars, and i^* denotes the rate of interest on securities denominated in foreign currency. The unexpected interest rate differential is denoted by $[(i - i^*)_t - E_{t-1}(i - i^*)_t]$. SE is the standard error of the equation. A quasi-R^2 was computed as $1 - \text{var}(\hat{u}_t)/\text{var}(\ln S_t)$.

TABLE 6

TWELVE-MONTH INTEREST RATE DIFFERENTIALS AND EXCHANGE RATES; INSTRUMENTAL VARIABLES
MONTHLY DATA, JUNE 1973–JULY 1979
(Standard Errors in Parentheses)

Dependent Variable $\ln S_t$	Constant	$\ln F_{t-1}$	$(i - i^*)_t$	$[(i - i^*)_t - E_{t-1}(i - i^*)_t]$	SE	R^2	D-W
Dollar/pound	024	.965	− 087	.978	.022	.97	1.82
	(.016)	(.021)	(.121)	(.286)			
Dollar/franc	−.180	.884	−.184	.915	.029	.79	2.16
	(.118)	(.181)	(.245)	(.432)			
Dollar/DM	−.013	.981	−.063	1.031	.031	.93	2.08
	(.053)	(.054)	(.338)	(.752)			

NOTE.—Interest rates are the 12-month Euromarket rates. The expected interest rate differential $E_{t-1}(i - i^*)_t$ was computed from a regression of the interest differential on a constant, two lagged values of the differential, and the logarithm of the lagged forward exchange rate. A two-stage least-squares estimation method was used. The instruments were a constant, two lagged values of the interest differential, Durbin's rank variable of the unexpected differential, and the logarithm of the lagged forward exchange rate. The actual interest rate differential is denoted by $(i - i^*)_t$, where i denotes the rate of interest on securities denominated in U.S. dollars, and i^* denotes the rate of interest on securities denominated in foreign currency. The unexpected interest rate differential is denoted by $[(i - i^*)_t - E_{t-1}(i - i^*)_t]$. SE is the standard error of the equation. A quasi-R^2 was computed as $1 - \text{var}(\hat{u}_t)/\text{var}(\ln S_t)$.

rates and the interest rate differential have been associated positively, thus indicating that during that inflationary period the same factors which induced a rise in the interest differential also induced a rise in the spot exchange rates. Furthermore, consistent with the hypothesis that current changes in exchange rates are primarily a response to new information, the evidence shows the importance of the innovations in the interest differential.

The principle that current exchange rates already reflect expectations concerning the future course of events implies that unanticipated changes in exchange rates are primarily due to innovations. Since the empirical work suggests that most of the actual changes in exchange rates are unanticipated, it follows that most of the actual changes in exchange rates are due to "news." In the present section this principle was applied to the analysis of the relationship between exchange rates and interest rate differentials. The principle, however, is general. For example, it implies that the relationship between a deficit in the balance of trade and the exchange rate depends crucially on whether the deficit was expected or not. A deficit that was expected may have no effect on the exchange rate, since the latter already reflected these expectations. In contrast, an unexpected deficit in the balance of trade may contain significant new information that is likely to be accompanied by large changes in the exchange rate.[26] This distinction might be useful in interpreting the weak and unstable relationship between the balance of trade and the exchange rate without having to rely on explanations like the *J* curve or on variable import and export elasticities.

IV. Exchange Rates and Prices

One of the striking facts concerning the relationship between prices and exchange rates during the 1970s has been the poor performance of the predictions of the simple versions of the purchasing power parity doctrine. The originators and proponents of the purchasing

[26] For a further elaboration on the relationship between exchange rates and the current account see Kouri (1976); Branson (1977); Branson, Halttunen, and Masson (1977); Dornbusch and Fischer (1980); and Rodríguez (1980). For a special emphasis on the role of innovations in the trade balance see Mussa (1979*b*), and for empirical evidence see Dornbusch (1980) and Hakkio (1980). It should be noted that the empirical work on the association between exchange rates and current account innovations faces some difficulties since, in contrast with data on financial variables like interest rates, data on the current account are not available at short intervals. Furthermore, findings on the association between exchange rates and current account innovations should be interpreted with care since, rather than reflecting the sensitivity of exchange rates to "news," they might just reflect invoicing practices according to which U.S. exports are invoiced in terms of U.S. dollars, while imports are invoiced in terms of foreign currencies.

power parity doctrine (Wheatley and Ricardo during the first part of the nineteenth century and Cassel during the 1920s) have viewed the doctrine as an extension of the quantity theory of money to the open economy. By now the consensus seems to be that, when applied to aggregate national price levels, purchasing power parities can be expected to hold in the long run if most of the shocks to the system are of a monetary origin which do not require changes in relative prices. To the extent that most of the shocks reflect "real" changes (like differential growth rates among sectors), the required changes in sectoral relative prices may result in a relatively loose connection between exchange rates and aggregate price levels. The experience during the 1970s illustrates the extent to which real shocks (oil embargo, supply shocks, commodity booms and shortages, shifts in the demand for money, differential productivity growth) result in systematic deviations from purchasing power parities. As illustrated in figures 1–3, short-run changes in exchange rates have not been closely linked to short-run differentials in the corresponding national inflation rates as measured by consumer price indices. Furthermore, this loose link seems to be cumulative. As illustrated in figures 4–6 divergences from purchasing power parities, measured in terms of the relationship between exchange rates and the ratio of consumer price indices, seem to persist.

The link between prices and exchange rates is illustrated in table 7, which reports the results of regressions of the various exchange rates on the corresponding ratios of wholesale and of cost-of-living price indices. As may be seen, the results of the regressions which involve the U.S. dollar are extremely poor. For the dollar/pound and the dollar/franc exchange rates, the estimates of the coefficients of the price ratios are insignificant and, for the dollar/DM exchange rate, the estimates differ significantly from unity. In contrast the results of the regressions of exchange rates that do not involve the U.S dollar or the U.S. price level (the pound/DM and the franc/DM exchange rates) are superior: Except for the wholesale price indices in the franc/DM regression all the coefficients are highly significant and the elasticities of the exchange rates with respect to the various price indices do not differ significantly from unity.

The vast difference in the performance of the regressions for the various currencies can be explained by noting that first, due to transport cost, purchasing power parities are expected to hold better among the neighboring European countries than among each of these countries and the United States; second, changes in commercial policies and nontariff barriers to trade seem to have been more stable within Europe than between Europe and the United States; third, within Europe the snake agreement and later on the European Mon-

TABLE 7

PURCHASING POWER PARITIES: INSTRUMENTAL VARIABLES MONTHLY DATA: JUNE 1973–JULY 1979
(Standard Errors in Parentheses)

Dependent Variable ln S_t	Constant	$\ln(P_w/P_w^*)$	$\ln(P_c/P_c^*)$	SE	D-W	ρ
Dollar/pound	.712	.165	...	027	1.63	.963
	(.149)	(.507)				
	2.982	...	1.070	.029	1.66	.998
	(2.978)		(.897)			
Dollar/franc	−1.521	.184	..	.029	2.26	863
	(.027)	(.374)				
	−1.570	...	−1.070	.029	2.30	901
	(.047)		(.817)			
Dollar/DM	−.900	1.786	...	.034	1.69	739
	(.018)	(.230)				
	−.908	...	2.217	.031	1.96	.759
	(.175)		(.263)			
Pound/DM	−1.668	.821	..	027	1.60	.895
	(.041)	(.144)				
	−1.666	...	.965	.027	1.57	.909
	(.048)		(.197)			
Franc/DM	.863	−.026	...	.020	1.61	.981
	(.143)	(.487)				
	602	..	1.180	.019	1.48	929
	(.048)		(.327)			

NOTE — The logarithm of the spot exchange rate is denoted by ln S_t. $\ln(P_w/P_w^*)$ and $\ln(P_c/P_c^*)$ denote, respectively, the logarithms of the ratios of the wholesale price indices and the cost of living indices. Cochrane-Orcutt iterative technique with a two-stage least-squares estimation method was used, the instruments are a constant, time, time squared, and lagged values of the dependent and independent variables. SE is the standard error of the equation, and ρ denotes the first-order autocorrelation coefficient.

etary System have resulted in a reduced degree of intra-European flexibility of exchange rates; and fourth, there seem to have been large changes in the equilibrium real exchange rate between the U.S. dollar and the European currencies.[27] It should be noted, however, that to some extent the overall poor performance of the purchasing power parities doctrine is specific to the 1970s. During the floating rates period of the 1920s, the doctrine seems to have been much more reliable.[28]

The preceding discussion accounted for the persisting deviations from purchasing power parities in terms of changes in real factors which affect equilibrium relative price structure. It should be noted, however, that even in the absence of such changes there is a presumption that, at least in the short run, exchange-rate fluctuations

[27] The failure of the regression of the franc/DM exchange rate on the ratio of the wholesale price indices is explained in terms of the large changes in the French intersectoral relative prices; for an elaboration see Frenkel (1981).

[28] For evidence see Frenkel (1976, 1978, 1980*a*) and Krugman (1978).

would not be matched by corresponding fluctuations of aggregate price levels. The discussion in Section II emphasized that in periods which are dominated by "news" which alters expectations, exchange rates (and other asset prices which are traded in organized markets) are expected to be highly volatile. Aggregate price indices, on the other hand, are not expected to reveal such a degree of volatility, since they reflect the prices of goods and services which are less durable and, therefore, are likely to be less sensitive to the "news" which alters expectations concerning the future course of events. It follows, therefore, that in periods during which there is ample "news" which causes large fluctuations in exchange rates, there will also be large deviations from purchasing power parities.[29]

The difference between the characteristics of exchange rates and national price levels is also reflected in their time-series properties and is fundamental for interpreting the deviations from purchasing power parities. The monthly changes in exchange rates exhibit little or no serial correlation, while national price levels do exhibit a degree of serial correlation. The "stickiness" exhibited by national price levels need not reflect any market imperfection but rather it may reflect the costs of price adjustment which result in the existence of nominal contracts of finite length. Likewise, it may reflect the results of confusion between relative and absolute prices and confusion between permanent and transitory changes. This difference between the time-series properties of exchange rates and prices is reflected in the low correlation between the practically random month-to-month exchange-rate changes and the serially correlated differences between national rates of inflation.

The different degrees of volatility of prices and exchange rates are illustrated in table 8, which reports the average absolute monthly percentage changes in the various exchange rates and prices. As is evident, the mean absolute change in the various spot exchange rates has been about 2 percent per month (and even slightly higher for the changes in the forward rate). The magnitudes of these changes have been more than double the magnitudes of the changes in most of the various price indices, as well as in the ratios of national price levels. For example, the mean monthly change in the cost-of-living price index was 0.4 percent in Germany, 0.7 percent in the United States,

[29] On this see Mussa (1979*a*). It is noteworthy that the emphasis in the text has been on the words "large fluctuations"; this should be contrasted with periods during which there are large secular changes in the exchange rate (like the changes which occurred during the German hyperinflation). During such periods the secular changes do not stem necessarily from "news" and need not be associated with deviations from purchasing power parities.

0.9 percent in France, and 1.2 percent in the United Kingdom. These differences are even more striking for the detrended series.

The notion that exchange rates have been volatile is clearly illustrated by figures 1–3 and by table 8. The comparison of the magnitudes of the changes in the exchange rates with the magnitudes of the changes in the price indices and in the ratios of national price levels may suggest, according to a narrow interpretation of the purchasing power parity doctrine, that exchange-rate fluctuations have been "excessive." The previous discussion, however, has emphasized that exchange rates, being the relative prices of assets, are fundamentally different from the price indices of goods and services and, therefore, are expected to exhibit a different degree of volatility in particular during periods that are dominated by "news." An alternative yardstick for measuring the degree of exchange-rate fluctuations would be a comparison with prices of other assets. Indeed, while exchange-rate changes have been large relative to changes in national price levels, they have been considerably smaller than changes in the prices of other assets like gold, silver, many other commodities that are traded in organized markets, and common stocks. For example, table 8 also reports the mean absolute monthly percentage change in stock-market indices. As may be seen, the mean monthly change in these indices ranged from over 3 percent in Germany to over 6 percent in the United Kingdom. By these standards it is difficult to argue that exchange rates have been excessively volatile.

Given the short-run deviations from purchasing power parities, it is relevant to explore whether these deviations tend to diminish with time or tend to persist or even grow in size. In order to examine the patterns of the deviations, the autocorrelation functions and the par-

TABLE 8

MEAN ABSOLUTE PERCENTAGE CHANGES IN PRICES AND EXCHANGE RATES
MONTHLY DATA: JUNE 1973–JULY 1979

	Variable					
				Exchange Rates against the Dollar		
Country	WPI	COL	Stock Market	Spot	Forward	COL/COL_{US}
U.S.	.009	.007	.037	...	.	.
U.K.	.014	.012	.066	.021	021	007
France	.011	009	.054	.020	.021	003
Germany	.004	.004	.030	.024	024	004

NOTE.—All variables represent the absolute values of monthly percentage changes in the data. WPI denotes the wholesale price index and COL denotes the cost-of-living index. Data on prices and exchange rates are from the IMF tape (May 1979 version). The stock market indices are from *Capital International Perspective* (monthly issues).

tial autocorrelation functions of these deviations for the wholesale and the cost-of-living price indices have been computed. The deviation from purchasing power parities during month t is denoted by Δ_t and is defined as:

$$\Delta_t = \ln S_t - \ln(P/P^*)_t. \tag{10}$$

Figures 11–13 illustrate the patterns of the deviations for the three exchange rates. As may be seen, the general pattern is very similar for the three exchange rates and for the two price indices. In all cases the autocorrelation function tails off at what seems to be an exponential rate, and in all cases the partial autocorrelation function shows a spike at the first lag. This pattern seems to indicate (as might have been expected on the basis of the time-series properties of exchange rates and price indices) that the deviations from purchasing power parities follow a first-order autoregressive process. It is noteworthy, however,

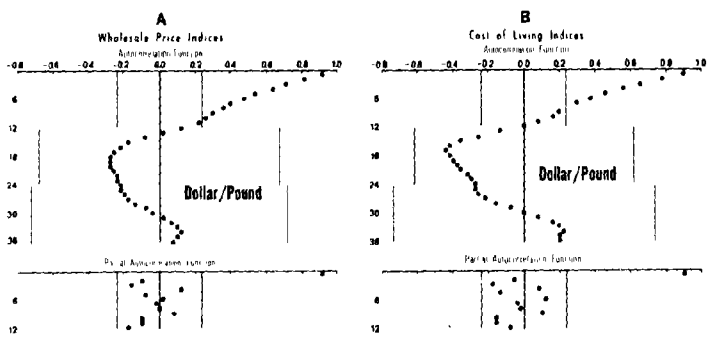


FIG. 11.—*A*, The dollar/pound: deviations from PPP with wholesale price indices. *B*, The dollar/pound: deviations from PPP with cost of living indices.

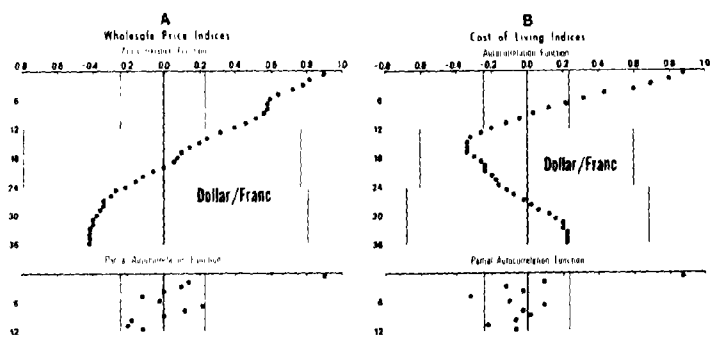


FIG. 12.—*A*, The dollar/franc: deviations from PPP with wholesale price indices. *B*, The dollar/franc: deviations from PPP with cost-of-living indices.

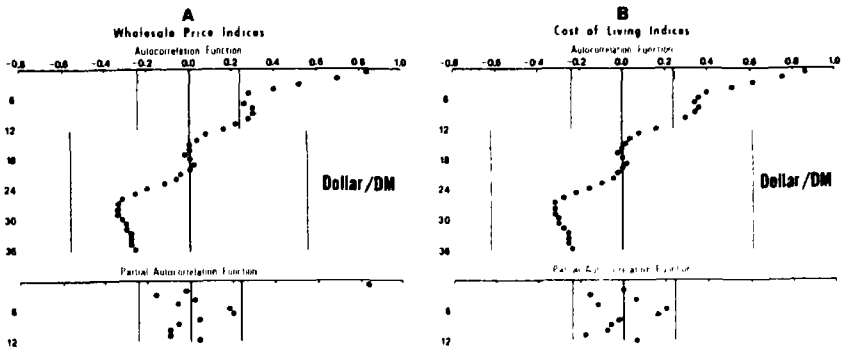


FIG. 13.—*A*, The dollar/DM: deviations from PPP with wholesale price indices. *B*, The dollar/DM: deviations from PPP with cost-of-living indices

that in all cases the value of the autoregression term is about .9, indicating the possibility that the series may not satisfy the stationarity requirement. To allow for this possibility, the autocorrelation functions and the partial autocorrelation functions of $\Delta_t - \Delta_{t-1}$, that is, of the first difference of the deviations from purchasing power parities, have also been examined. The results indicate that these differences are serially uncorrelated and thus imply that the deviations Δ_t follow a random-walk process. In view of this possibility it is concluded that the deviations from purchasing power parities seem to follow a first-order, autoregressive process, but that the data do not provide sufficient evidence to reject the alternative hypothesis of a random walk.[30]

V. Concluding Remarks

This paper examined some aspects of the operation of flexible exchange rates. The analysis was based on the experience of the 1970s. The principal conclusions which may be drawn from the empirical work are:

[30] If the deviations follow a random-walk process, then they do not entail (ex ante) unexploited profit opportunities. For a study of the deviations from purchasing power parities see Roll (1979). For an analysis of equilibrium deviations from purchasing power parities see Saidi (1977). It may be noted that the main difference between accepting the AR(1) rather than the random-walk hypothesis relates to the economic interpretation of the two alternative processes. The random-walk process implies that deviations from purchasing power parities do not tend to diminish with the passage of time, while the stable AR(1) process implies that there are mechanisms which operate to ensure that, in the long run, purchasing power parities are satisfied. For the purpose of forecasting the near future, however, there is very little difference between using the AR(1) process with an autoregressive coefficient of 0.9 and using the random-walk process.

i) In spite of the extraordinary turbulence in the markets for foreign exchange, it seems that, to a large extent, the markets have operated efficiently. It should be emphasized, however, that the concept of "efficiency" used in this context is somewhat narrow in that it only refers to the notion that the markets do not seem to entail unexploited profit opportunities. A broader perspective should deal with the social cost of volatility in terms of the interference with the efficiency of the price system in guiding resource allocation as well as with the cost of alternative outlets for the disturbances that are currently reflected in the volatility of exchange rates. As for the choice among alternative outlets for the disturbances, one may argue that since the foreign-exchange market is a market in which risk can be bought and sold relatively easily, it may be reasonable to concentrate the disturbances in this market rather than transfer them to other markets, such as labor markets, where they cannot be dealt with in as efficient a manner.

ii) The high volatility of exchange rates (spot and forward) reflects an intrinsic characteristic of the relative price of monies and other assets that are traded in organized exchange. The price of gold, the price of stocks, as well as exchange rates between national monies depend critically on expectations concerning the future course of events and adjust rapidly in response to new information. In this perspective, the exchange rate (in contrast with the relative price of national outputs) is being viewed as a financial variable which is determined in a macroeconomic setting.

iii) During inflationary periods, variations in nominal rates of interest are dominated by changes in inflationary expectations; as a result, high nominal rates of interest are associated with high exchange rates (a depreciated currency). This relationship was supported by the empirical work.

iv) The asset view of exchange-rate determination implies that "news" is among the major factors which influence changes in exchange rates. In this context the key finding was the dependence of exchange-rate changes on the unexpected changes in the rates of interest. This finding is in accord with the analytical prediction that current exchange rates already reflect current expectations about the future, while changes in the current exchange rates reflect primarily

changes in these expectations which, by definition, arise from new information.

v) The experience of the 1970s does not support the predictions of the simple version of the purchasing power parity doctrine which relates the values of current measured prices to current exchange rates. The empirical work showed that deviations from purchasing power parities can be characterized by a first-order autoregressive process.

vi) One of the key analytical insights that is provided by the monetary (or the asset-market) approach to the exchange rate is that exchange rates do not reflect only current circumstances but also reflect those circumstances which are expected to prevail in the future. This anticipatory feature of the exchange rate (which is emphasized by Mussa [1979*b*]) does not characterize (at least to such a degree) the prices of national outputs, which reflect to a large extent present and past circumstances as they are embedded in existing contracts. Consequently, periods which are dominated by large and frequent changes in expectations are likely to be periods in which the future is expected to differ greatly from the present and the past. Under such circumstances one may expect to find frequent deviations from purchasing power parities when the latter are computed using current prices. These deviations reflect the intrinsic difference between asset prices and national price indices.[31]

vii) Since commodity prices do not adjust fully in response to exogenous shocks, it seems that intervention in the foreign-exchange market, which ensures that exchange rates conform with purchasing power parities, would be a mistaken course of policy. When commodity prices are slow to adjust to current and expected economic conditions, it may be desirable to allow excessive adjustment in some other prices. Further, changes in real economic conditions requiring adjustment in the equilibrium relative prices of different national outputs occur continuously. An intervention rule which links changes in exchange rates rigidly to changes in

[31] It is interesting to note that this phenomenon was recognized by Gustav Cassel—he most recognized proponent of the purchasing power parity doctrine—according to whom: "The international valuation of the currency will, then generally show a tendency to anticipate events, so to speak, and become more an expression of the internal value that the currency is expected to possess in a few months, or perhaps in a year's time" (Cassel 1930, pp. 149–50).

domestic and foreign prices in accord with purchasing power parity ignores the occasional need for equilibrating changes in relative prices.

Appendix

1. *Exchange Rates*

The spot exchange rates are end-of-month rates obtained from the IMF tape (May 1979 version, updated to July 1979 using the November 1979 issue of the *International Financial Statistics*) obtained from the International Monetary Fund.

The forward exchange rates are end-of-month rates for 1-month maturity. The forward rates for the U.K. pound and the DM for the period June 1973–June 1978 are bid prices obtained from the International Money Market (IMM). For the period July 1978–July 1979 they are sell prices obtained from the *Wall Street Journal*. The forward rates for the French franc for the period June 1973–July 1974 are bid prices calculated from the *Weekly Review* publication of the Harris Bank which reports the spot rate and the forward premium; in each case the closest Friday to the end of the month was chosen. For the period August 1974–June 1978 the rates are bid rates obtained from the IMM, and for the period July 1978–July 1979 they are sell prices obtained from the *Wall Street Journal*.

2. *Prices*

The wholesale and cost-of-living price indices are period averages obtained from the IMF tape, lines 63 and 64, respectively.

3. *Rates of Interest*

All interest rates are 1-month Eurocurrency rates obtained from the *Weekly Review* of the Harris Bank. In all cases the figures used correspond to the last Friday of each month

4. *Stock Markets*

The stock market indices correspond to the last trading day of the month. The sources are *Capital International Perspective* (Geneva, Switzerland, monthly issues).

References

Barro, Robert J. "Unanticipated Money Growth and Unemployment in the United States." *A.E.R.* 67 (March 1977): 101–15.

Bilson, John F. O. "Rational Expectations and the Exchange Rate." In Frenkel and Johnson, eds., 1978.

———. "The 'Speculative Efficiency' Hypothesis." *J. Bus.* 54 (July 1981): 435–51.

Black, Stanley W. "International Money Markets and Flexible Exchange Rates." Princeton Studies in International Finance no. 32, Princeton Univ., March 1973.

Branson, William H. "Assett Markets and Relative Prices in Exchange Rate Determination." *Sozialwissenschaftliche Annalen* 1 (1977): 69–89.

Branson, William H.; Halttunen, Hannu; and Masson, Paul. "Exchange Rates in the Short Run: The Dollar-Deutschemark Rate." *European Econ. Rev.* 10 (December 1977): 303–24.

Cassel, Gustav. *Money and Foreign Exchange after 1914*. London: Macmillan, 1930.

Clements, Kenneth W., and Frenkel, Jacob A. "Exchange Rates, Money, and Relative Prices: The Dollar-Pound in the 1920's." *J. Internat. Econ.* 10 (May 1980): 249–62.

Dornbusch, Rudiger. "Capital Mobility, Flexible Exchange Rates and Macroeconomic Equilibrium." In *Recent Issues in International Monetary Economics*, edited by Emil Claassen and Pascal Salin. Amsterdam: North-Holland, 1976. (*a*)

———. "The Theory of Flexible Exchange Rate Regimes and Macroeconomic Policy." *Scandinavian J. Econ.* 78 (May 1976): 255–75. (Reprinted in Frenkel and Johnson, eds., 1978.) (*b*)

———. "Expectations and Exchange Rate Dynamics." *J.P.E.* 84, no. 6 (December 1976): 1161–76. (*c*)

———. "Monetary Policy under Exchange Rate Flexibility." In *Managed Exchange-Rate Flexibility: The Recent Experience*. Federal Reserve Bank of Boston Conference Series no. 20. Boston: Federal Reserve Bank of Boston, 1978.

———. "Exchange Rate Economics: Where Do We Stand?" *Brookings Papers Econ. Activity*, no. 1 (1980), pp. 143–85.

Dornbusch, Rudiger, and Fischer, Stanley. "Exchange Rates and the Current Account." *A.E.R.* 70 (December 1980): 960–71.

Edwards, Sebastian. "A Simple Monetary Model of Exchange Rate Determination in the Short Run—Some Preliminary Results for the Peruvian Experience 1950–1954." Unpublished manuscript, Univ. Chicago, 1979.

Fair, Ray C. "A Multicountry Econometric Model." Nat. Bur. Econ. Res. Working Paper Series no. 414, 1979.

Fischer, Stanley, ed. *Rational Expectations and Economic Policy*. Chicago: Univ. Chicago Press, 1980.

Frankel, Jeffrey A. "On the Mark: The Theory of Floating Exchange Rates Based on Real Interest Differentials." *A.E.R.* 69 (September 1979): 610–22

Frenkel, Jacob A. "A Monetary Approach to the Exchange Rate: Doctrinal Aspects and Empirical Evidence." *Scandinavian J. Econ.* 78 (May 1976): 200–224. (Reprinted in Frenkel and Johnson, eds., 1978.)

———. "The Forward Exchange Rate, Expectations and the Demand for Money: The German Hyperinflation." *A.E.R.* 67 (September 1977): 653–70.

———. "Purchasing Power Parity: Doctrinal Perspective and Evidence from the 1920's." *J. Internat. Econ.* 8 (May 1978): 169–91.

———. "Further Evidence on Expectations and the Demand for Money during the German Hyperinflation." *J. Monetary Econ.* 5 (February 1979): 81–96.

———. "Exchange Rates, Prices and Money: Lessons from the 1920's." *A.E.R.* 70 (May 1980): 235–42. (*a*)

———. "The Forward Exchange Rate, Expectations, and the Demand for Money—the German Hyperinflation: Reply." *A.E.R.* 70 (September 1980): 771–75. (*b*)

———. "The Collapse of Purchasing Power Parities during the 1970's." *European Econ. Rev.* 16 (February 1981): 145–65.

Frenkel, Jacob A., and Clements, Kenneth W. "Exchange Rates in the 1920's: A Monetary Approach." In *Development in an Inflationary World*, edited by M. June Flanders and Assaf Razin. New York: Academic Press, 1981.

Frenkel, Jacob A., and Johnson, Harry G., eds. *The Economics of Exchange Rates: Selected Studies.* Reading, Mass.: Addison-Wesley, 1978.

Frenkel, Jacob A., and Levich, Richard M. "Transaction Costs and Interest Arbitrage: Tranquil versus Turbulent Periods." *J.P.E.* 85, no. 6 (December 1977): 1209–26.

Frenkel, Jacob A., and Mussa, Michael L. "The Efficiency of Foreign Exchange Markets and Measures of Turbulence." *A.E.R.* 70 (May 1980): 374–81.

Frenkel, Jacob A., and Razin, Assaf. "Stochastic Prices and Tests of Efficiency of Foreign Exchange Markets." *Econ. Letters* 6, no. 2 (1980): 165–70.

Hakkio, Craig S. "Expectations and the Foreign Exchange Market." Ph.D. dissertation, Univ. Chicago, 1979.

———. "Exchange Rates and the Balance of Trade." Unpublished manuscript, Northwestern Univ., 1980.

Hansen, Lars Peter, and Hodrick, Robert J. "Forward Exchange Rates as Optimal Predictors of Future Spot Rates: An Econometric Analysis." *J.P.E.* 88, no. 5 (October 1980): 829–53.

Hausman, Jerry A. "Specification Tests in Econometrics." *Econometrica* 46 (November 1978): 1251–71.

Hodrick, Robert J. "An Empirical Analysis of the Monetary Approach to the Determination of the Exchange Rate." In Frenkel and Johnson, eds., 1978.

Isard, Peter. "Expected and Unexpected Changes in Exchange Rates: The Role of Relative Price Levels, Balance-of-Payments Factors, Interest Rates and Risk." International Finance Discussion Paper no. 156, Federal Reserve Board, April 1980.

Kouri, Pentti J. K. "The Exchange Rate and the Balance of Payments in the Short Run and in the Long Run: A Monetary Approach." *Scandinavian J. Econ* 78 (May 1976): 280–304.

Krugman, Paul R. "The Efficiency of the Forward Exchange Market: Evidence from the Twenties and the Seventies." Unpublished manuscript, Yale Univ., 1977.

———. "Purchasing Power Parity and Exchange Rates: Another Look at the Evidence." *J. Internat. Econ.* 8 (August 1978): 397–407.

Levich, Richard M. "Further Results on the Efficiency of Markets for Foreign Exchange." In *Managed Exchange Rate Flexibility: The Recent Experience*, Federal Reserve Bank of Boston Conference Series no. 20. Boston: Federal Reserve Bank, 1978.

———. "The Efficiency of Markets for Foreign Exchange." In *International Economic Policy: Theory and Evidence*, edited by Rudiger Dornbusch and Jacob A. Frenkel. Baltimore: Johns Hopkins Univ. Press, 1979.

Longworth, David. "An Empirical Efficient-Markets Model of Exchange Rate Determination." Unpublished manuscript, Bank of Canada, March 1980.

McCallum, Bennett T. "Topics concerning the Formulation, Estimation, and Use of Macroeconometric Models with Rational Expectations." Unpublished manuscript, Univ. Virginia, 1979.

McKinnon, Ronald I. "Floating Foreign Exchange Rates 1973–74: The Emperor's New Clothes." In *Institutional Arrangements and the Inflation Problem*,

edited by Karl Brunner and Allan H. Meltzer. Carnegie-Rochester Conference Series on Public Policy, vol. 3 (supplementary series to *J. Monetary Econ.*, 1976).

Mudd, Douglas R. "Do Rising U.S. Interest Rates Imply a Stronger Dollar?" *Federal Reserve Bank St. Louis Rev.* 61 (June 1979): 9–13.

Mussa, Michael L. "The Exchange Rate, the Balance of Payments and Monetary and Fiscal Policy under a Regime of Controlled Floating." *Scandinavian J. Econ.* 78 (May 1976): 229–48. (Reprinted in Frenkel and Johnson, eds., 1978.) (*a*)

———. "Our Recent Experience with Fixed and Flexible Exchange Rates: A Comment." In *Institutional Arrangements and the Inflation Problem*, edited by Karl Brunner and Allan H. Meltzer. Carnegie-Rochester Conference Series on Public Policy, vol. 3 (supplementary series to *J. Monetary Econ.*, 1976). (*b*)

———. "Exchange Rate Uncertainty: Causes, Consequences, and Policy Implications." Unpublished manuscript, Univ. Chicago, 1977.

———. "Empirical Regularities in the Behavior of Exchange Rates and Theories of the Foreign Exchange Market." In *Policies for Employment, Prices, and Exchange Rates*, edited by Karl Brunner and Allan H. Meltzer. Carnegie-Rochester Conference Series on Public Policy, vol. 11 (supplementary series to *J. Monetary Econ.*, 1979). (*a*)

———. "Anticipatory Adjustment of a Floating Exchange Rate." Unpublished manuscript, Univ. Chicago, 1979. (*b*)

———. "The Role of the Trade Balance in Exchange Rate Dynamics." Unpublished manuscript, Univ. Chicago, 1979. (*c*)

Nelson, Charles R. "Rational Expectations and the Estimation of Econometric Models." *Internat. Econ. Rev.* 16 (October 1975): 555–61.

Obstfeld, Maurice. "Expectations and Efficiency in the Foreign Exchange Market." Unpublished manuscript, Massachusetts Inst. Tech., 1978.

Rodríguez, Carlos Alfredo. "The Role of Trade Flows in Exchange Rate Determination: A Rational Expectations Approach." *J.P.E.* 88, no. 6 (December 1980): 1148–58.

Roll, Richard. "Violations of Purchasing Power Parity and Their Implications for Efficient International Commodity Markets." In *International Finance and Trade*, edited by Marshall Sarnat and Giorgio P. Szego. Vol. 1. Cambridge, Mass.: Ballinger, 1979.

Saidi, Nasser. "Rational Expectations, Purchasing Power Parity and the Business Cycle." Unpublished manuscript, Univ. Chicago, 1977.

Shiller, Robert J. "Do Stock Prices Move too Much to Be Justified by Subsequent Changes in Dividends?" Unpublished manuscript, Univ. Pennsylvania, 1979.

Singleton, Kenneth J. "Expectations Models of the Term Structure and Implied Variance Bounds." *J.P.E.* 88, no. 6 (December 1980): 1159–76.

The Role of Intergenerational Transfers in Aggregate Capital Accumulation

Laurence J. Kotlikoff

Yale University

Lawrence H. Summers

Massachusetts Institute of Technology

This paper uses historical U.S. data to directly estimate the contribution of intergenerational transfers to aggregate capital accumulation. The evidence presented indicates that intergenerational transfers account for the vast majority of aggregate U.S. capital formation; only a negligible fraction of actual capital accumulation can be traced to life-cycle or "hump" savings. A major difference between this study and previous investigations of this issue is the use of more accurate longitudinal age-earnings and age-consumption profiles. These profiles are simply too flat to generate substantial life-cycle savings. This paper suggests the importance of and need for substantially greater research and data collection on intergenerational transfers. Life-cycle models of savings that emphasize savings for retirement as the dominant form of capital accumulation should give way to models that illuminate the determinants of intergenerational transfers.

During the past quarter century, the life-cycle hypothesis (Modigliani and Brumberg 1954; Ando and Modigliani 1963) has emerged

We are indebted to R. Barro, M. Boskin, G. Chamberlain, R. Clower, M. Darby, S. Engerman, M. Feldstein, V. Fuchs, R. Gordon, E. Leamer, P. Menchik, J. Skinner, and F. Welch for enlightening discussions. We thank the Foundation for Research in Economic Education; the Department of Health, Education, and Welfare; and the National Bureau of Economic Research for providing research support. All views expressed here are solely those of the authors.

[*Journal of Political Economy*, 1981, vol. 89, no. 4]
© 1981 by The University of Chicago 0022-3808/81/8904-0005$01.50

as the principal theory of aggregate saving in the American economy. It has stimulated a vast quantity of research testing and exploring its implications. Despite this effort, the importance of lifetime consumption smoothing to the process of capital accumulation remains unresolved.[1] This paper presents evidence that the pure life-cycle component of aggregate U.S. savings is very small. American capital accumulation results primarily from intergenerational transfers. Distinguishing the roles of life-cycle savings and intergenerational transfers in the capital accumulation process is crucial to a number of economic issues. Economic models that incorporate intergenerational transfers can generate strikingly different results for a number of major economic questions, including the burden of the national debt, the impact of social security on savings, the incidence of taxation, the perpetuation of the inequality of wealth, and the optimal structuring of taxes to promote economic growth (see, e.g., Diamond 1970; Atkinson 1971; Feldstein 1974; Oulton 1976; Calvo, Kotlikoff, and Rodríguez 1979).

This paper has two objectives. The first is simply to answer an accounting question, namely, can life-cycle savings alone account for the U.S. capital stock? The second goal is to answer an economic question: If, ceteris paribus, there were no intergenerational transfers, how large would the U.S. capital stock be? That is, if all such transfers were taxed in a confiscatory way, by how much would capital accumulation be reduced? We find that life-cycle savings cannot explain the capital stock in an accounting sense and that in the absence of intergenerational transfers the U.S. capital stock would be substantially smaller.

Our findings are based on a new methodological approach.[2] In particular, a variety of historical U.S. data detailing population, labor earnings, consumption, and government taxes and transfers are used to directly estimate the shapes of historic age-earnings and age-consumption profiles. These profiles are combined with data on rates of return to calculate a stock of life-cycle wealth. This stock of life-cycle wealth is compared with aggregate wealth holdings in the United States. If there were no intergenerational transfers, the stock of life-cycle wealth would exactly equal total U.S. wealth. When intergenerational transfers occur, these two stocks differ by an amount equal to the stock of net received transfers. Hence, comparing total wealth with life-cycle wealth indicates whether the life-cycle model

[1] Tobin (1967) and Boskin (1978) argue that life-cycle savings are the predominant form of savings in the United States. White (1978) and Darby (1979) argue otherwise. Tobin and Dalde (1971) provide a somewhat intermediate view.

[2] Our methodological approach was influenced by and is similar to those of Brittain (1978) and Darby (1979).

alone can explain aggregate U.S. capital formation. We find that lifetime consumption and earnings profiles do not exhibit the kind of shapes needed to generate substantial life-cycle savings. Log linear approximations to these profiles suggest that growth rates of real earnings slightly exceed growth rates of real consumption over the lifetime. Since the life-cycle theory strongly relies on a lifetime growth rate of consumption in excess of the growth rate of earnings, the life-cycle theory of savings with no intergenerational transfers is a very poor description of the process of capital accumulation in the U.S. economy.

Our findings are in agreement with recent studies by White (1978) and Darby (1979) and rationalize other stylized facts about the U.S. economy at odds with the life-cycle theory. Mirer presents evidence from social security data that "the aged do not run down their wealth during their lifetime" (1979, p. 442). Indeed, after he adjusts for intercohort differences in wealth at retirement, Mirer finds "that wealth clearly increases with age" (1979, p. 442). Darby points out that, although the ratio of expected retirement years to expected life span increased by 67 percent from 1890 to 1930, aggregate savings rates showed no increase during this period as would be predicted by the life-cycle theory (1979, pp. 22–28). Atkinson (1971) and Oulton (1976) construct life-cycle models to determine how much of observed British inequality of wealth may be explained by this theory. The answer is, very little. After taking into account inequality in age-earnings profiles and realized rates of return, Oulton concludes, "The results indicate that none of these factors, either singly or in combination, are capable of accounting for a substantial proportion of actual wealth inequality" (1976, p. 99).

Section I presents a theoretical framework for considering the importance of intergenerational transfers to aggregate capital accumulation. Section II discusses the procedure to estimate the stock of life-cycle wealth. The data are described in Section III. Section IV presents and intuitively motivates the findings. The sensitivity of the findings to reasonable possible errors in the data and estimation procedure is also considered. The fifth section attempts to account for the residual between total U.S. wealth and the stock of life-cycle wealth using the limited information available on actual intergenerational transfers. The sixth and final section of the paper presents the conclusions and discusses some of their implications.

I. Life-Cycle and Transfer Wealth—a Theoretical Framework

The division of the stock of wealth, W, into life-cycle and transfer components, L and T, respectively, is easily understood by consider-

ing a 3-period model of economic growth with identical individuals in each age cohort. At a point in time, aggregate wealth in the economy is the sum of individual wealth holdings. Let W_i stand for the wealth holdings of a representative individual of age i, and let P_i stand for the population of individuals age i; then

$$W = W_1P_1 + W_2P_2 + W_3P_3. \tag{1}$$

The W_i's equal, by definition, accumulated flows of net received transfers plus accumulated flows of earnings net of government taxes less government transfers less accumulated flows of consumption, where accumulation takes place at the after-tax interest rate r. Let T^k_{ij} be the net transfer received at age i from individuals age j for individuals who are currently age k. For example, T^3_{12} represents the net transfer (which may be positive or negative) that a current 3-period-old individual received at the time he or she was 1 period old from individuals who were then 2 periods old. Let e^k_i and c^k_i represent the earnings and consumption of k-period-old individuals when they were i periods old. Assuming that earnings, consumption, and transfers occur at the end of each period, W_1, the wealth of age 1 individuals, equals zero. With this notation, W_2 and W_3 may be expressed as:

$$\begin{aligned} W_2 &= (T^2_{12} + T^2_{13})(1 + r) + (e^2_1 - c^2_1)(1 + r) \\ W_3 &= (T^3_{12} + T^3_{13})(1 + r)^2 + (T^3_{21} + T^3_{22})(1 + r) \\ &\quad + (e^3_1 - c^3_1)(1 + r)^2 + (e^3_2 - c^3_2)(1 + r). \end{aligned} \tag{2}$$

Combining (1) and (2) yields:

$$W = T + L, \tag{3}$$

where $T = (T^2_{12} + T^2_{13})(1 + r)P_2 + [(T^3_{12} + T^3_{13})(1 + r)^2 + (T^3_{21} + T^3_{23})(1 + r)]P_3$ and $L = (e^2_1 - c^2_1)(1 + r)P_2 + [(e^3_1 - c^3_1)(1 + r)^2 + (1 + r)(e^3_2 - c^3_2)]P_3$.

Equation (3) is the fundamental accounting relationship analyzed in this paper. Transfer wealth, T, corresponds to accumulated net received transfers, while life-cycle wealth, L, corresponds to accumulated earnings less accumulated consumption. Clearly, in a world with no intergenerational transfers, T would equal zero and W would equal L. Our first goal is to establish the relative magnitudes of the two components T and L and thereby determine whether U.S. wealth holdings can be predominantly explained by life-cycle savings. Since substantially less information is available about the values of the T^k_{ij}, most of this paper is devoted to calculating the value of L. Section V does, however, attempt to estimate T directly using fragmentary data and invoking steady-state assumptions.

If the economy is in a steady state, net intergenerational transfers

received at a given age are constant through time, so that:

$$\begin{aligned} T_{12} &= T_{12}^2 = T_{12}^3 \\ T_{13} &= T_{13}^2 = T_{13}^3 \\ T_{21} &= T_{21}^3 \\ T_{23} &= T_{23}^3. \end{aligned} \tag{4}$$

Under the assumption that population grows at a constant rate n,

$$P_1 = P_3(1 + n)^2 \quad \text{and} \quad P_1 = P_2(1 + n). \tag{5}$$

Since transfers received by age group i from age group j equal the negative of group j's transfers from age group i,

$$T_{ij}P_i = -T_{ji}P_j, \tag{6}$$

and

$$T_{12}P_1 = -T_{21}P_2. \tag{7}$$

From (4), (5), (6), and (7), T may be expressed in the steady state as:[3]

$$\begin{aligned} T = P_1\Bigg\{ T_{13}\left[\frac{1 + r}{1 + n} + \frac{(1 + r)^2}{(1 + n)^2}\right] \\ + T_{12}\frac{(1 + r)^2}{(1 + n)^2} + T_{23}\frac{(1 + r)}{(1 + n)^2}\Bigg\}. \end{aligned} \tag{8}$$

Under the assumption that $r = n$, T equals the yearly net flow of transfers from old to young cohorts, t, multiplied by the transfer-weighted age gap, g, between donors and recipients:

$$T = [P_1(T_{12} + T_{13}) + P_2T_{23}] \cdot \left[\frac{T_{12}P_1 + 2T_{13}P_1 + T_{23}P_2}{P_1(T_{12} + T_{13}) + P_2T_{23}}\right] \equiv t \cdot g. \tag{9}$$

If r exceeds (is less than) n, accumulated transfer wealth will exceed (be less than) the annual flow of transfers times the weighted average age gap. Equations (8) and (9) show that the contribution of transfers to the total stock of wealth depends critically on both the volume of the annual flow of transfers and the age span of transfers. Equation (8) is invoked in Section V to directly estimate the size of T.

The second goal of this paper is to ask the economic question, How would the U.S. capital stock, W, change if, because of changes in taxes or tastes, intergenerational transfer wealth, T, was reduced? Since consumption, leisure, and transfer decisions are jointly determined it is important to examine the possibility that changes in transfers might

[3] By examining (8), one can see that T corresponds simply to the accumulated holdings of transfers received by people who are currently alive from people who are currently dead.

induce or be associated with changes in consumption and earnings paths and thus alter the life-cycle component of wealth. In addition to recognizing interdependencies between transfer wealth and life-cycle wealth, the effect of changes in transfer behavior on capital formation may be examined in both partial and general equilibrium contexts. The partial equilibrium change in W resulting from a reduction in T holds wage rates and interest rates constant and corresponds to a shift in the household supply curve of capital. This paper considers only the steady-state partial equilibrium impact of changes in transfers on the capital stock. A calculation of the general equilibrium effect requires taking account of the responsiveness of the level and shapes of earnings, consumption, and lifetime transfer profiles to changes in capital accumulation.[4]

The Steady-State Partial Equilibrium Reduction in Capital Intensity Arising from a Reduction in Intergenerational Transfers

In order to analyze the partial equilibrium reduction in the stock of wealth, we first note that life-cycle wealth, L, equals accumulated earnings minus accumulated consumption. Accumulated earnings, in turn, equals accumulated wages at full-time work minus the accumulated value of leisure. Letting C stand for accumulated consumption, S for accumulated full-time wages, and M for the accumulated value of leisure, equation (3) becomes:

$$W = T + S - M - C. \tag{3'}$$

The partial equilibrium percentage reduction in W for a percentage change in transfer wealth, T, equals:

$$\frac{\partial W}{\partial T}\frac{T}{W} = \left(1 - \frac{\partial M}{\partial T} - \frac{\partial C}{\partial T}\right)\frac{T}{W}.^{5} \tag{10}$$

Equation (10) states that the proportionate reduction in the capital stock equals the share of transfer wealth in aggregate wealth, T/W, minus two additional terms indicating how accumulated earnings and accumulated consumption respond to changes in transfers.

Our analysis of the response terms $\partial M/\partial T$ and $\partial C/\partial T$ assumes that the utility of consumption and leisure is separable from the utility derived from intergenerational transfers. This assumption implies that the marginal rates of substitution between consumption and leisure at different points in time are independent of the level of

[4] Kotlikoff (1979) demonstrates that general equilibrium changes in capital intensity can be substantially smaller than partial equilibrium changes.

[5] Because wage rates are held constant in this partial equilibrium analysis, $\partial S/\partial T$ equals zero.

intergenerational transfers. Two examples of utility functions exhibiting this property are:

$$\begin{aligned} U_0 = \log C_0 + \log C_1 + \log C_2 + \log l_1 \\ + \log l_2 + \log l_3 + \alpha \log T_{12} + \alpha \log T_{13} \\ + \alpha \log T_{21} + \alpha \log T_{23} + \alpha \log T_{31} + \alpha \log T_{32} \end{aligned} \tag{11}$$

and

$$\begin{aligned} U_0 = \log C_0 + \log C_1 + \log C_2 + \log l_1 \\ + \log l_2 + \log l_3 + \alpha U_1. \end{aligned} \tag{11'}$$

The term U_0 stands for the utility of a representative individual of generation zero. The terms l_1, l_2, and l_3 correspond to leisure in different periods. In (11) the individual derives utility directly from the level of net transfers. In (11′) the individual derives utility from the utility, U_1, of his descendants. This is an example of an "overlapping utility function." For both types of separable utility functions the marginal rates of substitution between consumption and leisure are independent of the level of α, the preference parameter influencing the size of transfers. In addition, the first-order conditions for the optimal choice of consumption and leisure involve the equality between these marginal rates of substitution and the relative prices of consumption and leisure at different ages. Hence, neither changes in transfer preferences nor taxes on intergenerational transfers that affect only the price of transfers but not the price of consumption or leisure will alter the first-order conditions.

Given the levels of transfers, the utility-maximizing levels of consumption and leisure can be separately derived from these first-order conditions and the lifetime budget constraint given below:

$$\begin{aligned} C_1 + \frac{C_2}{1+r} + \frac{C_3}{(1+r)^2} + M_1 + \frac{M_2}{1+r} + \frac{M_3}{(1+r)^2} = s_1 \\ + \frac{s_2}{1+r} + \frac{s_3}{(1+r)^2} + T_{12} + T_{13} + \frac{T_{21}}{1+r} + \frac{T_{23}}{1+r} \\ + \frac{T_{31}}{(1+r)^2} + \frac{T_{32}}{(1+r)^2}. \end{aligned} \tag{12}$$

In equation (11), M_1, M_2, and M_3 are the values of leisure in periods 1, 2, and 3 (by value of leisure we mean the number of units of leisure times the price [wage] per unit). The terms s_1, s_2, and s_3 are full-time wages in periods 1, 2, and 3. Equation (12) indicates that lifetime consumption and leisure are financed by full-time lifetime earnings plus net lifetime received transfers. The separability assumption im-

plies that changes in preferences or taxes that alter the values of the T_{ij} have only an income effect on the choices of consumption and leisure over the lifetime.

We now demonstrate that this income effect of a change in transfers is identically zero for the case that r equals n.[6] Using (5) and (6) to rewrite (12), we obtain:

$$C_1 + \frac{C_2}{1+r} + \frac{C_3}{(1+r)^2} + M_1 + \frac{M_2}{1+r} + \frac{M_3}{(1+r)^2} = s_1 + \frac{s_2}{1+r} + \frac{s_3}{(1+r)^2} + \frac{T_{12}(r-n)}{(1+r)} + \frac{T_{23}(r-n)}{(1+r)^2} + \frac{T_{13}[(1+r)^2 - (1+n)^2]}{(1+r)^2}. \tag{12'}$$

We can also write the budget constraint (12′) in terms of aggregate transfer wealth per young person T' where $T' = T/P_1$:

$$C_1 + \frac{C_2}{1+r} + \frac{C_3}{(1+r)^2} + M_1 + \frac{M_2}{1+r} + \frac{M_3}{(1+r)^2} = s_1 + \frac{s_2}{1+r} + \frac{s_3}{(1+r)^2} + \frac{(r-n)(1+n)^2 T'}{(1+r)^3}. \tag{12''}$$

As is clear from (12′) and (12″), when $r = n$, lifetime consumption and leisure are financed solely out of lifetime full earnings; all of the accrued interest on received transfers, as well as the principle itself, is used to maintain steady-state transfers per head at a constant level. When $r = n$, reducing transfers has no impact on the steady-state budget constraint, and the terms $\partial M/\partial T$ and $\partial C/\partial T$ in (10) are zero (i.e., since lifetime consumption and earnings paths stay the same when $r = n$, steady-state life-cycle wealth, L, will be unaltered by changes in steady-state transfer wealth, T). Thus, any decline in transfer wealth, T, will reduce total wealth, W, dollar for dollar in partial equilibrium when $r = n$.

In the case that r differs from n, steady-state changes in the level and pattern of transfers (the T_{ij} terms) affect the steady-state budget constraint only insofar as they alter the stock of transfer wealth T. This is a general proposition that can easily be shown to hold independent of the number of periods in the model. When r exceeds n, part of lifetime consumption and leisure is financed by lifetime transfers. Again, under the separability assumption, the reduction in trans-

[6] In the case of a tax on transfers, the income effect of a change in transfers when $r = n$ will be zero only for the case of a compensated tax on transfers. Compensated rather than uncompensated tax changes are the appropriate focus of studies of government tax policies toward savings. See Diamond 1970 and Kotlikoff and Summers 1979.

fers has simply an income effect on consumption and leisure and, assuming both are normal goods, will reduce lifetime consumption and raise lifetime earnings. In (10) both $\partial M/\partial T$ and $\partial C/\partial T$ will be positive when r exceeds n. The reverse will be true when n exceeds r. It is easy to demonstrate that introducing labor-augmenting productivity growth simply requires relabeling n everywhere as the population growth rate plus the productivity growth rate. In the case of productivity growth, the steady state is characterized by a constant level of transfers per effective worker.

The value of these terms depends on the particular preferences determining the levels and shapes of consumption and leisure paths. As one example, table 1 gives the value of the transfer wealth-adjustment factor $[1 - (\partial M/\partial T) - (\partial C/\partial T)]$ for a particular specification of preferences for consumption and leisure and for different values of r and n. The example assumes a 55-year life span and uses the logarithmic utility function with time-preference parameter ρ and leisure-preference parameter α:[7]

$$U = \int_0^D \log C_t e^{-\rho t} dt + \alpha \int_0^D \log l_t e^{-\rho t} dt. \tag{13}$$

The l_t terms in (13) are units of leisure at different points in time. The calculations underlying table 1 are presented in Kotlikoff and Summers (1980, appendix A).

The figures in table 1 indicate that the response of consumption and labor supply can be quite important in determining the final partial equilibrium effect of a reduction in transfers when the interest rate differs from the economy's growth rate. For the specific logarithmic function chosen here, a real interest rate that is 1 percent higher than the rate of population plus productivity growth generates total wealth response numbers approximately equal to 0.7. If, on the other hand, the rate of population plus productivity growth exceeds the real interest rate by 1 percent, reducing transfer wealth by $1.00 reduces aggregate wealth by $1.40.

For the United States the annual rate of population growth has averaged 1.40 percent from 1900 through 1974, the period of this study (see U.S. Department of Commerce 1975, pt. 1, p. 8). Productivity growth as measured by the annual percentage change in real GNP per man-hour has averaged 2.20 percent over this period.[8]

[7] The utility function in (13) corresponds only to that part of total utility pertaining to consumption and leisure. Any set of preferences for transfers may be added to (13) as long as utility of consumption and leisure is separable from the utility of transfers. The assumption of a steady state with productivity growth also requires that the entire utility function be homothetic.

[8] U.S. Department of Commerce 1973, series A167 and A168, p. 210, and U.S. Department of Labor 1978, p. 336. Kendrick's series of output per man-hour is used for the years 1900–1910, and the BLS series thereafter.

TABLE 1

TRANSFER WEALTH ADJUSTMENT FACTOR

r	n	ρ	$1 - \frac{\partial M}{\partial T} - \frac{\partial C}{\partial T}$
.02	.01	02	.73
.02	.01	03	71
.02	.02	.02	1.00
.02	.03	.02	1 40
.02	.04	.02	2.00
.03	.01	.01	.58
.03	01	.05	.48
.03	.02	.03	.71
.03	.03	03	1.00
.03	.04	.03	1.43
04	.01	.01	.45
.04	.01	07	.30
.04	.02	.04	49
.04	03	.04	70
04	.04	04	1.00

Adding 1.4 to 2.2 yields 3.6 percent as the rate of population plus productivity growth. This paper calculates a portfolio-weighted net nominal rate of return for the U.S. economy from 1900 through 1974. After subtracting inflation, the real annual net after-tax rate of return in the U.S. economy averaged 4.5 percent between 1900 and 1974.[9]

The fact that real interest rates have on average exceeded growth rates suggests that part of U.S. consumption and leisure may well have been financed by interest earned on intergenerational transfers. Hence, a dollar reduction in transfer wealth can be expected to reduce aggregate wealth by less than $1.00. The calculated 1 percent gap between r and n suggests that eliminating transfer wealth in the U.S. economy would reduce total wealth by about 70 percent of the amount of transfer wealth. While the 70 percent figure is meant to be suggestive rather than precise, it appears that almost any choice of preferences would yield offset factors not far from 70 percent.[10] That

[9] To obtain a real net rate of return series, we subtract the annual percentage change in the CPI from our net nominal return series indicated in U.S. Department of Commerce (1975, pt. 1, pp. 210–11, table B2) and U.S. Department of Commerce (1978, p. 490).

[10] These calculations were performed for the more general isoelastic utility function. A wide range of parameter values for rates of risk aversion and time preference produced values of $[1 - (\partial M/\partial T) - (\partial C/\partial T)]$ ranging from 0.6 to 0 75 when r exceeds n by 1 percent. While it is unclear what utility function is most appropriate, these parameter values cover a wide range of different shapes of consumption and leisure profiles and, we presume, bound the likely response. The assumption implicit in these utility functions that consumption and leisure respond homothetically to changes in

is, the life-cycle offset to a change in transfer wealth is likely to be small.

II. The Estimation of Life-Cycle Wealth—Methodological Approach

Total life-cycle wealth in the U.S. economy equals the sum over all living persons in the economy of life-cycle assets. Each person's life-cycle assets correspond to his or her accumulated earnings less accumulated consumption where accumulation occurs at the actually realized rates of return. If data existed detailing each person's earnings, consumption, and realized rate of return on assets at each point in time in the past, it would be easy to check whether U.S. wealth holdings could be explained predominantly by life-cycle accumulation.

Obviously, such detailed individual specific data are not available. However, historical data for the United States on aggregate earnings, aggregate consumption, rates of return, and age-earnings and age-consumption profiles may be used to carry out this life-cycle asset computation on a cohort-by-cohort basis. This paper treats individuals of each sex within an age cohort as identical and estimates the average excess of after-tax earnings plus government transfers over consumption experienced by members of that age-sex cohort during each of their adult years in the past. These differences are then accumulated up to 1974 using historical net nominal interest rates. The total over all age-sex cohorts of these accumulated life-cycle assets is then compared with the 1974 value of total U.S. private net worth.

Life-cycle wealth of the age-sex cohort that is age a and sex j in 1974, $L_j(a)$, is given by

$$L_j(a) \equiv \sum_{i=18}^{a} \{[\bar{E}_j(a, i) + \bar{G}_j(a, i) - \bar{C}_j(a, i)] \prod_{k=i}^{a} [1 + r(a, k)]\} P_j(a). \tag{14}$$

transfers may be inappropriate. Even if the utility function is homothetic, capital market constraints precluding borrowing when young against future earnings may make the consumption and leisure response to a reduction in transfers nonhomothetic. These capital market constraints may not be binding in the presence of transfers because the excess of consumption over earnings when young is financed by the received transfers. Eliminating or reducing transfers could make these capital market constraints binding and require proportionately less consumption and leisure when young than when old.

In equation (13), $P_j(a)$ stands for the number of people alive in 1974 who are age a and sex j. The terms $\bar{E}_j(a, i)$, $\bar{G}_j(a, i)$, and $\bar{C}_j(a, i)$ are, respectively, the average after-tax earnings, government transfers, and consumption of the age a–sex j cohort at the time its members were age i. Age 18 is taken as the age of adulthood. Consumption expenditures by adults on children under the age of 18 are considered to be a part of adult consumption rather than intergenerational transfers.[11] The term $r(a, k)$ is the economy-wide annual net nominal interest rate received by the age a cohort during the year the cohort's members were k years old.

Our estimate of total life-cycle assets in the economy in 1974 equals the sum over all age-sex cohorts of the estimated values of $L_j(a)$.[12] The difference between life-cycle wealth and total U.S. wealth in 1974 is our estimate of transfer wealth.

Calculation of Longitudinal Profiles of Net Earnings, Government Transfers, and Consumption

The essential idea involved in these calculations is to use cross-sectional distribution profiles to allocate aggregate flows of net earnings, government transfers, and consumption to different age-sex cohorts in a given year. By performing this computation for each year from 1900 to 1974, we obtain the longitudinal profiles $\bar{E}_j(a, i)$, $\bar{G}_j(a, i)$, and $\bar{C}_j(a, i)$ as i varies.

To illustrate the computation for $\bar{E}_j(a, i)$, we define the following terms: $\bar{e}_t$ is the average earnings of 40-year-old male workers in year t; $g_m(a, t)$ is the ratio of average earnings of male workers at age a in year t to average earnings of 40-year-old male workers in year t; $g_f(a, t)$ is the ratio of average earnings of female workers at age a in year t to average earnings of 40-year-old female workers in year t; λ_t is the ratio of average earnings of 40-year-old female workers to average earnings of 40-year-old male workers in year t; $\alpha_m(a, t)$ is the percentage of males age a with work experience in year t; $\alpha_f(a, t)$ is the percentage of females age a with work experience in year t; H_t is total after-tax labor income in year t; $P_m(a, t)$ is the population of males age a in year t; and $P_f(a, t)$ is the population of females age a in year t. Given information on all of the above variables except $\bar{e}_t$, equation (14) is used to solve for $\bar{e}_t$:

[11] Treating consumption of children under age 18 as their own life-cycle consumption would greatly reduce our estimate of life-cycle wealth and greatly increase our estimate of transfer wealth. The age of adulthood is somewhat arbitrary, but 18 appears to be a reasonable number and of general interest.

[12] Assets of children under age 18 are assumed to be completely inherited and are not included in life-cycle wealth.

$$H_t = \bar{e}_t \sum_{a=18}^{100} [g_m(a, t)\alpha_m(a, t)P_m(a, t) + \lambda_t g_f(a, t)\alpha_f(a, t)P_f(a, t)]. \tag{15}$$

Given $\bar{e}_t$, $\bar{E}_m(a, i)$ for males and $\bar{E}_f(a, i)$ for females satisfy:

$$\bar{E}_m(a, t) = \bar{E}_m[a, a - (1974 - t)] = \bar{e}_t g_m(a, t)\alpha_m(a, t), \tag{16}$$

$$\bar{E}_f(a, t) = \bar{E}_f[a, a - (1974 - t)] = \bar{e}_t \lambda_t g_f(a, t)\alpha_f(a, t). \tag{17}$$

The procedure for computing the longitudinal profiles of consumption, $\bar{C}_j(a, i)$, is identical to that just described for earnings; a cross-sectional profile of relative consumption by age and sex is used to distribute aggregate U.S. consumption in each year to different age-sex cohorts.

Calculation of Net Nominal Interest Rates Series

The interest rate term $r(a, t)$ on (13) depends only on the year t since $r(a, t) = r[1974 - (a - t)] = r(t)$. The computation is performed using two different interest rate time series. Series 1 was calculated by dividing U.S. wealth holdings into six separate assets, estimating a rate of return series for each asset, and then weighting each asset's rate of return by its share in the U.S. portfolio for the particular year in question.

Series 2 is based on the wealth augmentation relation:

$$W_{t+1} = W_t(1 + r_t) + E_t + G_t - C_t. \tag{18}$$

Equation (18) indicates that private U.S. net worth in year $t + 1$ equals U.S. net worth in year t plus savings in year t. Savings in year t is, in turn, equal to net income on assets, $r_t W_t$, plus after-tax labor income and government transfers, E_t and G_t, less aggregate consumption in year t, C_t. Equation (18) is solved for r_t using time-series data on W_t, E_t, G_t, and C_t.

III. Description of Data Used to Calculate Life-Cycle Wealth

This section describes sources and procedures for several of our data series. Kotlikoff and Summers (1980) provide a detailed description of all the data and adjustments to the data used in calculations.

Information on compensation of employees after 1929 is obtained from the *National Income Accounts* and, for the years 1909–29, from Leven, Moulton, and Warburton (1934) and Kuznets (1941, 1946). Prior to 1909, employee compensation is imputed from Kendrick's estimates of net national product using the ratios of employee com-

pensation to national product from 1909 to 1918.[13] The imputation of labor income to the self-employed follows Christensen (1971). The number of proprietors is multiplied by the average earnings of full-time equivalent employees. The chief bias here is simply that the self-employed may earn more or less on average than employees. To insure that the results do not reflect a substantial underestimate of entrepreneurial earnings, the estimate of entrepreneurial earnings for each year is increased by 20 percent.

Estimates of state and federal income taxes paid on labor income were obtained from IRS *Statistics of Income* and the *National Income and Product Accounts*. Combined employer and employee social security and health insurance taxes were obtained from the *Social Security Annual Statistical Supplement* for various years.

Estimates of the age-earnings profiles for men and women were obtained by fitting separate regressions to social security estimates of median annual earnings of workers at different ages for the years 1950 through 1975.[14] The general shapes of the profiles predicted by the regressions are quite similar throughout the 1950–75 period. For the years prior to 1950, the predicted male and female age-earnings profiles for the year 1955 were used.[15]

The value of λ_t, the ratio of average earnings of 40-year-old female workers to average earnings of 40-year-old male workers in year t, was taken to be .55 throughout the period. This appears to be an upper-bound value for λ_t (see Kotlikoff and Summers 1980, p. 46, n. 35).

Values for work experience rates by age and sex, $\alpha_m(a, t)$ and $\alpha_f(a, t)$, are available only after 1959. Substantially more information, especially for the early 1900s, is available on labor force participation rates by age and sex. Regression analysis for the post-1959 period indicates that work experience rates can be predicted quite well by functions of age and labor force participation rates. This post-1959 regression was used to estimate the α function for each year from 1900 to 1974 based on labor force participation rates for the appropriate year. The labor force participation rates equal the values predicted from regressions of labor force participation rates on fifth-order age polynomials for each sex and for different census years.

While cross-sectional distribution functions by sex and age were computed for social security and medicare transfers, we assumed that

[13] See Leven et al. 1934, p. 155, Kuznets 1941, pp. 322–23, and Kendrick's series A6 and B61 in U.S. Department of Commerce (1973), pp. 182, 222. Because of limitations on population estimates on individuals above age 85, we treat all individuals who were above age 85 in 1974 as if they were age 85. Since 85-year-olds in 1974 did not reach the age 18 until 1907, we really do not use any data for the years prior to 1907 in this paper.

[14] These and other regression results are reported in Kotlikoff and Summers (1980).

[15] The year 1955 was chosen because by 1955 the social security data cover most of the private economy's work force.

other transfers, including veterans' benefits and welfare payments, were distributed in the cross-section over the years 1900–1974 according to that year's age-earnings distributions.

Total consumption expenditure, Z_t, is taken from National Income Accounts after 1929 and from Kendrick prior to 1929. Ideally one should subtract expenditures on consumer durables and add imputed rent on consumer durables to obtain true economic consumption. The difficulties of implementing this for the pre-1929 years led us to simply use the consumer expenditure series. In any case, there appears to be very little difference between the consumer expenditure series and the true economic consumption series.[16]

The cross-sectional consumption distribution functions were obtained from the 1960 and 1972–73 *Consumer Expenditure Survey* (CES) tapes. Unfortunately, similar data are not available to generate these profiles for earlier years. Life-cycle wealth is computed using the 1972–73 cross-sectional profiles throughout the period 1900–1974. The 1960 profiles generated essentially the same level of life-cycle wealth.

In distributing total household consumption to household members, household heads and their spouses are assumed to consume equally; all other household members, including children, were allocated 50 percent of the household head's consumption. The total consumption of children under the age of 18 was then reallocated to the household head and spouse, assigning each one half of children's consumption in the case of two spouses, or giving all the children's consumption to the household head if he or she was single. The general shape of the profiles was quite insensitive to the assumption that other household members and children consume only 50 percent of the consumption level of the household head. The 50 percent assumption generated slightly more life-cycle wealth and is the one used.

Net nominal interest rate series 1 is generated from data on historical rates of return and Goldsmith and Lipsey's (1963) data on portfolio shares. From Goldsmith and Lipsey's balance sheets we constructed seven asset categories plus liabilities. These are tangible, noncorporate business assets including land and structures, residential land and structures, money, short-term claims (savings accounts and U.S. Treasury bills), corporate stock, long-term corporate bonds, and U.S. savings bonds. A rate of return series was calculated for each asset as well as the liabilities. A weighted rate of return was computed

[16] Christensen and Jorgenson's (1973) "true" consumption series (p. 17) is very close to the *NIA* series for the years 1929–69. The Christensen and Jorgenson series is slightly higher for most years; using the Christensen and Jorgenson series for the post-1929 years would, therefore, lead us to calculate less life-cycle wealth.

taking the share of each item in net worth during the period considered as the weight. The nominal rates of return were converted to after-tax nominal rates of return using average income tax rates for each year. The series 2 net nominal interest rates were computed for the post-1929 period using equation (22). Reliable net worth data are not available until 1929. In the computation the series 2 rates from the pre-1929 years are filled in with the series 1 values.

IV. The Size of Life-Cycle Wealth

At the end of 1974 total net worth in the U.S. economy equaled $4.154 trillion. Of the $4.154 trillion, $134 billion represent the tangible assets of nonprofit institutions (see FRB-MIT balance sheets); in 1974 nonprofit institutions owned 4.4 percent of corporate equities (SEC 1977, p. 11). Under the assumption that this percentage applies to all other financial assets, total net worth of nonprofit institutions in 1974 would equal $270 billion.[17] After subtracting $270 billion from $4.154 trillion, the 1974 net worth figure for the noninstitutional household sector is $3.884 trillion. This is the number to compare with the life-cycle wealth calculations.

In addition to presenting total life-cycle wealth of individuals, a second life-cycle wealth concept, LCW2, is calculated that adds to LCW1 an upper-bound estimate of accumulated interspousal transfers. Many economists would include within their definition of life-cycle wealth the amount of accumulation that arises from interspousal transfers. Interspousal transfers give rise to wealth accumulation, first, because wives are on average younger than husbands and, second, because females live longer than males. Since it is conceptually difficult, if not impossible, to exactly trace individual patterns of household formation that involve marriage, divorce, and death, an upper-bound estimate of the stock of interspousal accumulated transfers is added to LCW1. To be specific, at age 21 all males are assumed to marry 18-year-old females.[18] If either spouse dies prior to reaching

[17] See SEC 1977, p. 11. This assumes that nonprofit institutions have no liabilities.

[18] The 1972–73 CES data indicate that the average age gap between husband and wife is 3 years. The number of deaths in an age-sex cohort in a particular year was obtained by multiplying age-, sex-, and year-specific death rates from the U.S. census bureau's population estimates. *Historical Statistics* (U.S. Department of Commerce 1975, vol. 1, death rates, series B 181–92, pp. 61–62) are smoothed. Even in the absence of marriage, our LCW2 concept would be of interest because it essentially treats individuals as if their wealth at each moment were fully annuitized; i.e., given the uncertainty of the date of death, individuals in a life-cycle model with no bequest motive would always purchase annuities. Hence, our procedure of essentially passing a cohort's bequest over to surviving members mimics the operation of an annuities market. Thus LCW2 can be thought of as essentially the life-cycle wealth of the U.S. economy if individuals always purchased annuities.

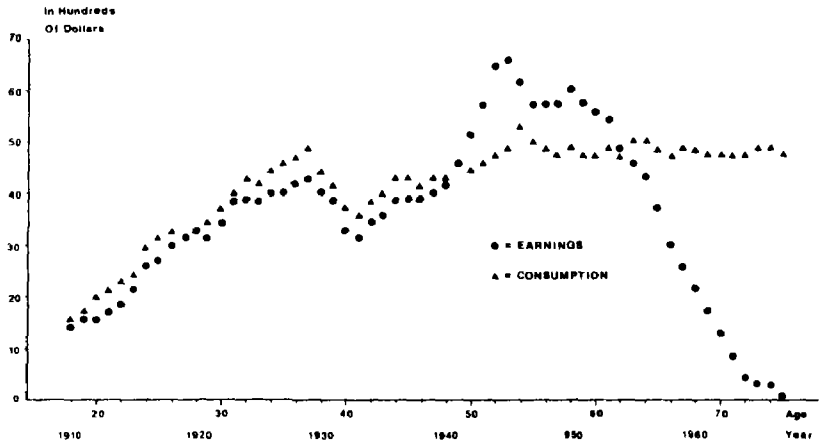


FIG. 1.—Sum of male and female longitudinal average earnings and average consumption profiles, age 18 in 1910–age 82 in 1974.

when old. The historical reality for the United States is simply that longitudinal earnings and consumption profiles have not exhibited the kinds of shapes required for substantial life-cycle accumulation. For males, earnings profiles greatly exceed consumption profiles over most of the life cycle. For females, however, the opposite is true. Life-cycle wealth starts out negative and remains negative over all ages. Figures 1 and 2 graph longitudinal profiles of the sum of male and female average earnings plus government transfers and average consumption for two different age cohorts. The profiles are presented in real 1967 dollars. The diagrams clearly show that the male excess of earnings over consumption is essentially offset by the female excess of consumption over earnings.

Contrary to the life-cycle simulation studies that have generated substantial life-cycle wealth, the actual growth rate of lifetime consumption does not substantially exceed the actual growth rate in lifetime earnings.[20] For example, the male age cohort that reached

[20] Roger Gordon has suggested to us that observed age-earnings profiles may differ from true age-earnings profiles to the extent that firms and workers are engaged in implicit contractual arrangements. Firms could pay out to workers less (more) than they truly earn when young and more (less) than they truly earn when old. Some part of a worker's life-cycle saving or dissaving would then be accomplished within the firm and would correspond to a claim (or liability) attached to the firm's assets. While we would strongly contest the empirical validity of this proposition, certainly a very upper-bound estimate of life-cycle savings with firms would be the difference between the market and replacement costs of capital in the corporate sector. Using 0.819, the average value of q over the period 1952–74 (von Furstenberg 1977) and the 1974 replacement value of corporate capital of $1.679 trillion gives an upper-bound estimate of $304 billion for this effect.

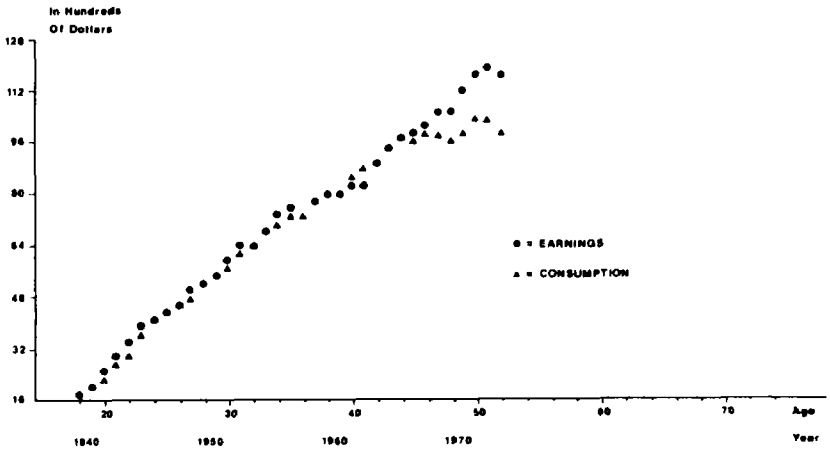


FIG. 2.—Sum of male and female longitudinal average earnings and average consumption profiles, age 18 in 1940–age 52 in 1974.

age 18 in 1920 experienced a growth rate of 2.93 percent in real earnings between ages 18 and 65. In comparison, the male growth in real consumption between ages 18 and 72 was only 2.32 percent. For the corresponding female cohort, earnings growth was 2.23 percent and consumption growth only 1.72 percent. Real consumption profiles are fairly flat, and real earnings profiles peak in late middle age, not at early ages. Both of these factors militate against the life-cycle theory of savings as an explanation of U.S. wealth holdings.

The large transfer wealth value of $3.151 trillion cannot simply be explained as private transfers from old to young offsetting forced government social security and medicare transfers from young to old. Under the Barro (1974) view, introducing unfunded social security into an economy will not lead to a change in consumption but will increase private net transfers from old to young. If Barro's view is correct, it is easy to determine how large private transfers would have been with no social security and medicare system by simply setting all historic social security and medicare benefits and taxes to zero in the calculation. This procedure led to a series 1 value of $775 billion and a series 2 value of $1,094 billion for LCW2. Hence, even if Barro is correct in his view about unfunded social security, transfer wealth would still have exceeded $2.79 trillion in 1974 if there had never been a social security program.

One possible problem with the estimates is that the age-consumption profiles prior to 1960 may have looked substantially different from those after 1960. To explore this issue, relative consumption prior to age 40 was reduced by 10 percent and relative

consumption after age 40 was increased by 10 percent. Thus, 60-year-olds were effectively assigned an extra 20 percent consumption relative to 30-year-olds in every year from 1900 to 1974. Altering consumption profiles in this manner produced values of −193 for LCW1 and 1,702 for LCW2 using series 1 and −169 for LCW1 and 2,178 for LCW2 using series 2. These numbers are still substantially smaller than the $3,884 billion of 1974 U.S. net worth.

Another issue of concern is whether our data series on aggregate flows accurately describe U.S. experience in the 1900s. As mentioned, to insure against underestimation of the labor income of the self-employed, we have already increased our estimates by 20 percent. Raising this add-on factor to 30 percent increases the series 1 values of LCW1 to −679 and LCW2 to 872. It is worth noting that eliminating any self-employment add-on factor lowers LCW1 to −1,739 and LCW2 to −714.

A final data issue concerns consumption expenditures of nonprofit institutions as well as contributions of individuals to such institutions. Ideally we should subtract institutional consumption from aggregate consumption and add contributions to these institutions to arrive at total household consumption. While not all the data needed for this adjustment are available, what data are available suggest that the adjustment would be very small and would slightly lower our estimate of life-cycle wealth.[21]

The LCW2 wealth concept effectively deals with possible bias arising from differential survival probabilities between the rich and the poor. This calculation assumes that at least one member of each household survives to age 75 and attributes all household life-cycle accumulation to the surviving spouse(s). Hence, this procedure assumes that rich and poor households have identical survival probabilities, which eliminates this issue of bias.

Another type of aggregation error could arise if some households continually received higher rates of return on their assets than other households. In order to investigate this issue one would need detailed knowledge of the joint distribution of rates of return and household consumption and earnings patterns. Unfortunately, this information cannot be obtained from existing micro data, and this fact precludes a reasonable assessment of the magnitude of this source of error.

[21] In 1970, for example, the level of philanthropic payments of individuals totaled $16.09 billion (U.S. Department of Commerce 1975, vol. 1, series H 399–400, p. 359). In comparison, the *NIA* (U.S. Department of Commerce 1976, p. 90, lines 101 and 102) reported 1970 expenditures of institutions for religious, educational, and welfare activities totaled only $11.32 billion.

V. Explaining the Residual—the Stock of Transfer Wealth

This section investigates the extent to which U.S. data on intergenerational transfers can explain the large residual between total net worth and the estimates of life-cycle wealth. Unfortunately, there are very few data detailing nontaxed intergenerational transfers. These nonenumerated transfers take many forms. Parents who lend money at below market rates to their children for a down payment on a house or a business are engaged in an intergenerational transfer. A father who makes his son a full partner in a lucrative business can effectively transfer large sums of money with no tax liability to the IRS. Parents who fully or partially support their children through college or after college are making transfers. A grandmother's gift of her expensive wedding china and rings to her granddaughter is an intergenerational transfer. Transfers in these forms, as well as outright monetary gifts, are very rarely reported.

To obtain some idea about how large aggregate intergenerational transfer flows need to be in order to explain a $3.151 trillion stock of transfer wealth, consider the multiperiod analogue to equation (8):

$$T = t\frac{e^{(r-n)D}}{(r-n)}[1 - e^{(n-r)(G-D)}]. \tag{19}$$

Equation (19) is a simplified expression for the steady-state stock of transfers; the formula assumes that all transfers are given at age G and received at age I. The yearly flow of intergenerational transfers is t, the age of death D, and r and n are rates of interest and population plus productivity growth, respectively. In the case $r = n$, this expression reduces to $T = t(G - I)$, the analogue to equation (9). Table 3 evaluates the age gap factor, that is, the terms multiplying t for various parameter values and taking D to be 55.[22]

Exactly which age gap is appropriate is unclear.[23] An age gap of 30 allows some significant transfers to grandchildren as well as children. The age gap factor in this case for a .01 excess of r over n is 45. Hence, to explain $3,151 billion in transfer wealth, the yearly flow of transfers would have to equal 3,151 divided by 45, or $70 billion. A 2 percent differential between $(r - n)$, again assuming a 30-year gap, would require only a $46 billion annual transfer flow.

The 1962 Federal Reserve Survey of Consumer Finances was used

[22] If we take age 18 as the age of adulthood, the value of 55 for D corresponds to a real-world age of death of 73.

[23] To precisely calculate a weighted age gap we would need data detailing age of donors and recipients as well as the size of the transfers. Such data are currently unavailable.

TABLE 3

AGE GAP FACTOR

	$r - n$					
AGE GAP $(G - I)$	−.02	−.01	0	.01	.02	.03
20	8	13	20	31	49	78
25	11	16	25	38	59	92
30	14	20	30	45	68	103
35	17	24	35	51	76	113

to estimate the bequest portion of the yearly transfer flow. Specifically, the distribution of net worth holdings by age, sex, and marital status was calculated. Next, 1962 mortality probabilities were applied to this distribution to arrive at an estimated distribution of bequests by age, sex, and marital status. An additional adjustment to these figures was made in order to reconcile the U.S. aggregate net worth value estimated from the Federal Reserve tape with the FRB-MIT net worth data.

Unpublished data collected by Paul Menchik and Martin David from the Washington, D.C., Inheritance Tax File for the year 1967 reveal that males who were married at the time of their death left 10 percent of their estate to their children. Married female decedents left 19 percent of their estates to their children. The proportion left to grandchildren and other young relatives is unclear, but it probably does not exceed 2 percent for males and 4 percent for females, which represents half of the percentage contribution to other relatives. For single male decedents, 32 percent of the estate is left to children; the single female proportion is 37 percent. Again taking half of the contributions to other relatives as "distant in age" intergenerational transfers, another 22 percent can be ascribed to single males and 23 percent to single females. Using these figures and the Federal Reserve simulated bequest distributions, the estimated "distant in age" intergenerational bequest transfer flow in 1962 was $11.9 billion. Multiplying this figure by the 221.6 percent growth in the nominal value of total U.S. net worth between 1962 and 1974 gives us an estimate of the 1974 bequest transfer flow of $26.4 billion.

In 1974 the total value of life insurance death benefits equaled $8.885 billion. In conjunction with the Menchik numbers, this raises the $26.4 billion bequest flow figure to $28.9 billion.

We also estimate the flow of intergenerational transfers from parents to children that occur in the form of financial support during college. In 1974 college enrollment totaled 8.8 million students (U.S.

Department of Commerce 1978, p. 138). In 1976 parental contributions to college-enrolled children who were taken as tax deductions averaged $1,738 (Froomkin 1977, p. 479). Assuming that non-tax dependent college-enrolled children received one-quarter of this level of support, average support from parents in 1976 was $1,270. Reducing this number by 15 percent, the growth in tuition between 1974 and 1976 (U.S. Department of Commerce 1978, p. 165) suggests a level of college support payments of about $1,080 in 1974. This type of intergenerational transfer could then account for $10.3 billion of the total 1974 flow. Adding the 10.3 to the 28.9 leaves us with $39.2 billion of explained transfer flows.

Another component of the intergenerational transfer flow is transfers made in the form of trusts. While direct data on the value of new trusts formed in 1974 are not available, there are fiduciary income tax data for the years 1965 and 1970 that permit a rough calculation of this transfer flow component (U.S. Internal Revenue Service 1970, p. 46). Between 1965 and 1970 the number of new trusts established each year averaged 35,098. The 1970 income of the 152,398 existing trusts in 1970 totaled $7,513 billion. The average 1970 trust had, therefore, an income of $9,985. Dividing this value by our 1970 series 1 interest rate of 0.0787 gives an estimate of the average value of a 1970 trust of $126,874. Multiplying this figure by 35,098 leads to an estimate of $4.44 billion as the value of new trusts established in 1974. After multiplying this number by 1.395 to allow for the growth in total wealth between 1970 and 1974, our estimate of the 1974 flow of intergenerational transfers in the form of trusts is $6.19 billion. This figure raises the explained transfer flow to $45.4 billion.

This $45.4 billion figure seems small compared with the $70 billion total flow needed if the stock of transfer wealth is $3.151 trillion and the age gap factor is 45. On the other hand, it is not too far from the $46 billion total flow needed if r exceeds n by 2 percent and the age gap factor is 68.

Unfortunately, before one can precisely determine the total intergenerational flow, substantially more information will have to be collected about family gift giving and support payments to children and grandchildren. These data may prove particularly difficult to obtain since we would have to estimate the value of the cedar chest passed from grandmother to grandnephew, or the family car given to son John as a college graduation present, or the value to son Alex of making him a full partner in a lucrative family business. Since the distribution of wealth is very highly skewed, such surveys need to be aimed at the intergenerational transfer payments of the very wealthy. However, the very wealthy may be the least willing to disclose these types of transfers because of potential estate and gift tax liabilities.

VI. Summary and Conclusions

The evidence presented in this paper rules out *life-cycle hump saving* as the major determinant of capital accumulation in the U.S. economy. Longitudinal age-earnings and age-consumption profiles do not exhibit the kinds of shapes needed to generate a large amount of life-cycle wealth accumulation. The view of U.S. capital formation as arising, in the main, from essentially homogeneous individuals or married spouses saving when young for their retirement is factually incorrect.

Intergenerational transfers appear to be the major element determining wealth accumulation in the United States. Our best estimates of the 1974 stock of transfer wealth after allowing for interspousal life-cycle accumulation is approximately $3 trillion.

While these estimates of the stock of transfer wealth are quite large, totally eliminating transfers in the U.S. economy would not necessarily reduce total U.S. wealth by the full amount of transfer wealth. We have demonstrated within the context of a steady-state growth model that a $1.00 reduction in the stock of transfer wealth may reduce total wealth by less than $1.00 if the steady-state real interest rate exceeds the steady-state growth rate. Taking the U.S. historical real interest and growth rates as illustrative, eliminating a $3 trillion stock of transfer wealth would reduce total U.S. wealth by about $2.1 trillion in a steady-state context. This, however, is a partial equilibrium analysis. Substantially more research must be undertaken before we can begin to attach probable numbers to full general equilibrium responses to changes in transfers.

This paper suggests the importance of and need for substantially more research and data collection on intergenerational transfers. Economic models of savings that stress the homogeneity of agents and the importance of the demographic structure should give way to models that emphasize the rather massive intergenerational transfers in the U.S. economy and the apparent concentration of these transfers among the very wealthy.

References

Ando, Albert, and Modigliani, Franco. "The 'Life Cycle' Hypothesis of Saving: Aggregate Implications and Tests." *A.E.R.* 53 (March 1963): 55–84.

Atkinson, A. B. "The Distribution of Wealth and the Individual Life-Cycle." *Oxford Econ. Papers*, n.s. 23 (July 1971): 239–54.

Barro, Robert J. "Are Government Bonds Net Wealth?" *J.P.E.* 82, no. 6 (November/December 1974): 1095–1117.

Boskin, Michael J. "Is Heavy Taxation of Capital Socially Desirable?" *Hearings before the Subcommittee on Economic Growth and Stabilization, the Joint Economic Committee, Congress of the United States, Ninety-fifth Congress*, July 1977. Washington: Government Printing Office, 1978.

Brittain, John A. *Inheritance and the Inequality of Material Wealth.* Washington: Brookings Inst., 1978.

Calvo, Guillermo A.; Kotlikoff, Laurence J.; and Rodríguez, Carlos Alfredo. "The Incidence of a Tax on Pure Rent: A New (?) Reason for an Old Answer." *J.P.E.* 87, no. 4 (August 1979): 869–74.

Christensen, Laurits R. "Entrepreneurial Income: How Does It Measure Up?" *A.E.R.* 61 (September 1971): 575–85.

Christensen, Laurits R., and Jorgenson, Dale W. *U.S. Income, Saving, and Wealth, 1929–1969.* Madison: Univ. Wisconsin, 1973.

Darby, Michael R. *Effects of Social Security on Income and the Capital Stock.* Washington: American Enterprise, 1979.

Diamond, Peter A. "Incidence of an Interest Income Tax." *J. Econ. Theory* 2 (September 1970): 211–24.

Feldstein, Martin S. "Social Security, Induced Retirement, and Aggregate Capital Accumulation." *J.P.E.* 82, no. 5 (September/October 1974): 905–26.

Froomkin, Joseph. *Tax Treatment of Tuition Expenses.* Testimony to the Committee on Ways and Means, Hearings before the Committee on Ways and Means, House of Representatives, Ninety-fifth Congress, Serial 95–56. Washington: Government Printing Office, 1977.

Goldsmith, Raymond W., and Lipsey, Robert E. *Studies in the National Balance Sheet of the United States.* Vol. 1. Princeton, N.J.: Princeton Univ. Press (for Nat. Bur. Econ. Res.), 1963.

Kotlikoff, Laurence J. "Social Security and Equilibrium Capital Intensity." *Q.J.E.* 93 (May 1979): 233–53.

Kotlikoff, Laurence J., and Summers, Lawrence H. "Tax Incidence in a Life Cycle Model with Variable Labor Supply." *Q.J.E.* 93 (November 1979): 705–18.

———. "The Role of Intergenerational Transfers in Aggregate Capital Accumulation." Working Paper no. 445, Nat. Bur. Econ. Res., February 1980.

Kuznets, Simon S. *National Income and Its Composition, 1919–1938, I and II.* New York: Nat. Bur. Econ. Res., 1941.

———. *National Product since 1869.* New York: Arno (for Nat. Bur. Econ. Res.), 1946.

Leven, Maurice; Moulton, Harold G.; and Warburton, Clark. *America's Capacity to Consume.* Washington: Brookings, 1934.

Mirer, Thad W. "The Wealth-Age Relation among the Aged." *A.E.R.* 69 (June 1979): 435–43.

Modigliani, Franco, and Brumberg, Richard. "Utility Analysis and the Consumption Function: An Interpretation of Cross-Section Data." In *Post-Keynesian Economics*, edited by Kenneth K. Kurihara. New Brunswick, N.J.: Rutgers Univ. Press, 1954.

Oulton, Nicholas. "Inheritance and the Distribution of Wealth." *Oxford Econ. Papers*, n.s. 28 (March 1976): 86–101.

Tobin, James. "Life Cycle Saving and Balanced Growth." In *Ten Economic Studies in the Tradition of Irving Fisher*, edited by William Fellner et al. New York: Wiley, 1967.

Tobin, James, and Dalde, Walter. "Wealth, Liquidity, and Consumption." In *Consumer Spending and Monetary Policy: The Linkages.* Federal Reserve Bank of Boston Monetary Conference, vol. 5. Boston: Federal Reserve Bank of Boston, 1971.

U.S. Department of Commerce, Bureau of the Census. *Historical Statistics of the U.S., Colonial Times to 1970.* Washington: Government Printing Office, 1975.

———. *U.S. Statistical Abstract.* Washington: Government Printing Office, 1978.

U.S. Department of Commerce, Bureau of Economic Analysis. *Long Term Economic Growth, 1860–1970.* Washington: Government Printing Office, 1973.

———. *The National Income and Product Accounts of the United States, 1929–1974, Statistical Tables.* Washington: Government Printing Office, 1976.

U.S. Department of Labor, Bureau of Labor Statistics. *Employment and Training Report of the President.* Washington: Government Printing Office, 1978.

U.S. Internal Revenue Service. *Statistics of Income.* Washington: Government Printing Office, 1970.

U.S. Securities and Exchange Commission. *Statistical Bulletin.* Washington: Government Printing Office, 1977.

von Furstenberg, George M. "Corporate Investment: Does Market Valuation Matter in the Aggregate?" *Brookings Papers Econ. Activity*, no. 2 (1977), pp. 347–97.

White, Betsy Buttrill. "Empirical Tests of the Life-Cycle Hypothesis." *A.E.R.* 68 (September 1978): 547–60.

Output Variability under Monetary Policy and Exchange Rate Rules

Warren E. Weber

Virginia Polytechnic Institute and State University

This paper examines the controversy concerning the relative desirability of fixed versus flexible exchange rates by examining whether a country can achieve a smaller variance of domestic output around its full employment path operating under an exchange rate rule or under a money supply rule. The paper finds that neither policy will always dominate the other. However, it does find that if shocks in one market of the economy are large relative to shocks to other markets, then one type of rule can be shown to dominate the other.

I. Introduction

One of the major concerns of the international trade literature has been whether or not a country is better off under a fixed or under a flexible exchange rate system. There is no consensus that one type of exchange rate scheme will always dominate the other. Rather, the optimal exchange rate scheme for a particular country has been found to depend upon the degree of openness of the country, the degree of international capital and labor mobility, and its rate of adjustment to disturbances.[1] Another consideration has been whether or not the disturbances experienced by the country are more likely to be external or internal in origin, although recently the emphasis has

I am indebted to Francis Ahking, Russell Boyer, Rodney Falvey, Anatoli Kuprianov, Leslie Lipschitz, Michael Salemi, Wilson Schmidt, members of the Workshop on Monetary Theory at VPI&SU, and two anonymous referees for useful comments. I take responsibility for any remaining errors.

[1] For a good survey of the conditions under which one exchange rate arrangement will dominate the other, see Tower and Willett (1976).

Journal of Political Economy, 1981, vol. 89, no. 4]
© 1981 by The University of Chicago. 0022-3808/81/8904-0007$01.50

shifted to the market in which the disturbance occurs rather than its geographical origin.[2]

This paper examines the controversy concerning the relative desirability of fixed versus flexible exchange rates from a different point of view. The question examined in this paper is whether a country can achieve a smaller variance of domestic output around its full employment path operating under an exchange rate rule or under a money supply rule. The exchange rate rules we consider will be completely general, so that a country operating under an exchange rate rule can be considered to be engaged in managed floating and, as a consequence, will have an endogenous domestic money supply. On the other hand, when an economy operates under a money supply rule, its exchange rate becomes endogenous, so that an economy operating under a money rule can be thought of as having perfectly floating exchange rates. The money supply rules we consider will also be completely general.

The model which underlies this analysis is an aggregate supply–aggregate demand model for an open economy with perfect capital mobility and rational expectations. The model is a stochastic and dynamic version of that used by Parkin (1977). It also has the same basic framework as Dornbusch (1976) but departs from his actual model by assuming continuous market clearing and rational expectations. The model is presented in Section II. The reduced-form solutions for real output under the alternative policy rules are found in Section III. In Section IV the variability of output under the two alternative policy rules is compared and some implications are presented. A summary is presented in Section V.

II. The Model

Following Dornbusch (1976) and Parkin (1977) we model a country which is small enough in the world capital market that it cannot affect the interest rate it faces. Thus, capital mobility will insure the equality of foreign and domestic interest rates after adjustment for expected exchange rate appreciation or depreciation. We also assume that domestic goods are imperfect substitutes for foreign goods, so that the domestic price level can differ from the foreign price level by more than the current exchange rate. Further, we assume that the country is small enough that its demands for imports will not affect the price of imports, and we arbitrarily set the price of foreign goods equal to unity.[3]

[2] Two recent articles which focus on disturbances by market origin rather than geographical origin are those by Boyer (1978) and Lipschitz (1978).

[3] By assuming that domestic output and foreign output are not perfect substitutes,

The basic model consists of the following four equations:

$$y_t = a_1(P_t - E_{t-1}P_t) + a_2 y_{t-1} + \mu_t \text{ (aggregate supply curve)}, \quad (1)$$

$$y_t = b_1[i_t - E_{t-1}(P_{t+1} - P_t)] + b_2(e_t - P_t) + \nu_t \text{ (IS curve)}, \quad (2)$$

$$m_t = c_0 P_t + c_1 y_t + c_2 i_t + c_3 e_t + \eta_t \text{ (LM curve)}, \quad (3)$$

$$i_t = \bar{i} + E_{t-1}(e_{t+1} - e_t) \text{ (interest rate parity)}. \quad (4)$$

The variables are defined as follows: y_t = the level of real output (income); m_t = the domestic nominal money supply; P_t = the price level; i_t = one plus the nominal interest rate; e_t = the exchange rate (the domestic currency price of foreign currency); $\bar{i}$ = one plus the world nominal interest rate; and μ_t, ν_t, η_t, ϵ_t, and ω_t = white-noise stochastic disturbance terms. All variables are expressed in logarithms. The operator E_{t-1} denotes the rationally formed expectation based upon information available at time $t - 1$. All stochastic error terms are assumed to be white noise and distributed independently of each other. Later in the analysis the operator σ_x^2 will be used to denote the variance of the random variable x.

The aggregate supply curve (1) is similar in form to that proposed by Lucas (1973) and has been used in a closed economy macro model by Sargent and Wallace (1975) and McCallum (1978) and in an open economy model by Flood (1979). It incorporates the natural rate hypothesis since real output is affected only by unanticipated changes in the price level. Following Gray (1976) and Fischer (1977*b*), we justify (1) as follows. Assume that nominal wage contracts for period t are negotiated at the end of period $t - 1$ and equate expected labor supply and expected labor demand conditional on the information available when the contracts are negotiated. Actual employment in period t, however, depends upon the realized wage in the period, and for this reason innovations in the price level cause fluctuations in employment and output.[4] Further, since employment varies inversely

our model differs from the models of Bilson (1978) and Flood (1979). We will also use money rules in our analysis which are more general than the random-walk money growth rules which they use.

[4] The aggregate supply curve (1) is also obtained by Gordon (1978, chap. 7) using assumptions very similar to those of Gray (1976) and Fischer (1977*b*). We have not used Lucas's derivation of (1) as he assumes informationally separated markets. As Flood has pointed out, the assumption of informationally separated markets on the supply side is not consistent with the contemporaneous trading across countries implicit in eqq. (2) through (4). Such contemporaneous trading also implies that the behavioral parameters in the model can be assumed to remain unchanged as the monetary authority switches from an exchange rate rule to a money rule or vice versa. As Lucas (1973), Barro (1976), and Leiderman (1979) have shown, such an assumption may not be correct if agents operate in informationally separated markets and must make inferences about the price level or the exchange rate based upon transactions prices ob-

with the real wage, we restrict $a_1 > 0$. The omission of a constant term in (1) indicates that y_t should be interpreted as the deviation of output about the level which would obtain if there had been no past innovations in the price level.

The IS curve (2) is that of conventional theory with two modifications. The first is that the interest rate variable is the real rather than the nominal interest rate. This modification is required by the fact that when the price level is allowed to change in the analysis, investment demand should depend upon real rather than nominal interest rates.[5] The second is the presence of the relative price of foreign to domestic goods. (The logarithm of the foreign price level is omitted since the foreign price level was assumed to be unity.) Since higher real interest rates should depress the demand for domestic output and since an increase in the price of foreign goods relative to domestic goods should increase the demand for domestic goods, we expect $b_1 \leqslant 0$ and $b_2 \geqslant 0$.

That the LM curve (3) is the conventional LM curve can be seen by rewriting (3) as

$$m_t - c_0 P_t - c_3 e_t = c_1 y_t + c_2 i_t + \eta_t. \tag{3'}$$

Thus, (3′) states that the demand for real balances depends upon real income and nominal interest rates and that for the money market to be in equilibrium the demand for real balances must be equal to the supply of real balances. The reason for including the exchange rate in the determination of real balances is that individuals in the economy may deflate their holdings of nominal money by a combination of both domestic and foreign prices since they can purchase both domestic and foreign goods.[6] We expect that $c_0, c_3 \geqslant 0$ and $c_0 + c_3 = 1$. Further, the conventional demand-for-money view is that $c_1 \geqslant 0$ and $c_2 \leqslant 0$.

Equation (4) results from the assumption of the smallness of the country in world capital markets and from the assumption that foreign and domestic assets are perfect substitutes for investors. As a consequence, capital mobility will force equality of expected nominal returns on domestic and foreign bonds expressed in units of the same currency.

In equations (2) and (4) of the model, expectations concerning the

served in their market. It would be interesting to examine how the conclusions of this paper would change if this informational assumption were modified.

[5] In general, for a random variable x, $\log(Ex) > E(\log x)$. Therefore, $i_t - E_{t-1}(P_{t+1} - P_t)$ is only approximately equal to the logarithm of one plus the real interest rate. Further, (4) is only an approximation to the interest rate parity condition for the same reason. However, $\log(Ex)$ is equal to the expected value of the first-order Taylor series expansion of $\log x$ around Ex. I am indebted to a referee for this point.

[6] This same assumption is made by Boyer (1978, p. 1047).

rates of change of the price level and the exchange rate are assumed to be conditioned on information available in period $t - 1$. On the other hand, period t information on the price level, the nominal interest rate, and the exchange rate are assumed to be used by agents in their output and money demand decisions. This raises the possibility that the model may be asymmetric with respect to its information assumptions, although such asymmetry could be reasonable if bond purchases were assumed to be made at the beginning of period t before the output and money demand decisions. Nonetheless, we recognize that there is a degree of arbitrariness with regard to the assumptions concerning the conditioning of expectations. For this reason we treat the case where agents use period t information to form expectations in the Appendix. Both assumptions about the conditioning of expectations have appeared previously in the literature. The papers by Sargent and Wallace (1975), McCallum (1978), and Flood (1979) use expectations conditioned on information available at time $t - 1$, whereas expectations conditioned on information available at time t are used by Bilson (1978) and by Flood in his appendix.

III. Implications

We now solve the model given by (1) through (4) first under a money policy rule and then under an exchange rate rule.

Money Rule

In an economy in which the monetary authority is following a rule for the determination of the money supply, the exchange rate is an endogenous variable; that is, the country operates under flexible exchange rates. When e_t is endogenous, solving equations (1) through (4) for the price level yields

$$\begin{aligned} P_t = k\{&(c_3 + b_2c_1)(a_1E_{t-1}P_t - a_2y_{t-1}) \\ &+ c_3b_1[\bar{\imath} + E_{t-1}(e_{t+1} - e_t) - E_{t-1}(P_{t+1} - P_t)] \\ &- b_2c_1[\bar{\imath} + E_{t-1}(e_{t+1} - e_t)] + b_2m_t + \theta_t\} \end{aligned} \tag{5}$$

where $\theta_t = [-(c_3 + c_1b_2)\mu_t + c_3\nu_t - b_2\eta_t]$ and $k = (b_2 + a_1c_3 + c_1a_1b_2)^{-1}$. Applying the E_{t-1} operator to both sides of (5) and subtracting the result from (5) yields

$$P_t - E_{t-1}P_t = k(\theta_t + b_2\epsilon_t), \tag{6}$$

where $\epsilon_t = m_t - E_{t-1}m_t$ is the innovation in the money supply in period

t. Substituting (6) into (1), we obtain the reduced-form solution for real output as

$$\begin{aligned} y_t &= a_1 k(\theta_t + b_2\epsilon_t) + a_2 y_{t-1} + \mu_t \\ &= k[b_2\mu_t + a_1 c_3 \nu_t + a_1 b_2(\epsilon_t - \eta_t)] + a_2 y_{t-1}. \end{aligned} \tag{7}$$

Given (7), the solution for the real exchange rate $(e_t - P_t)$ can be obtained from (2) and (3) using the Sargent-Wallace forward solution technique. We find[7]

$$\begin{aligned} e_t - P_t = &-b_1\bar{\imath}/b_2 + \{a_2/[b_2 + b_1(a_2 - 1)]\}y_{t-1} \\ &+ k[\mu_t - (1 + c_1 a_1)\nu_t + a_1(\epsilon_t - \eta_t)]. \end{aligned} \tag{8}$$

Substituting (7) and (8) into (2) yields the solution for the real interest rate as

$$i_t - E_{t-1}(P_{t+1} - P_t) = \bar{\imath} + \{a_2(a_2 - 1)/[b_2 + b_1(a_2 - 1)]\}y_{t-1}. \tag{9}$$

Further, from (7) we obtain the variance of real output under a money rule to be

$$\sigma^2_{y|M} = k^2(1 - a_2^2)^{-1}[b_2^2\sigma^2_\mu + (a_1c_3)^2\sigma^2_\nu + (a_1b_2)^2(\sigma^2_\eta + \sigma^2_\epsilon)]. \tag{10}$$

Since $E_{t-1}m_{t+j}$, $j \geq 0$, terms do not appear in the solutions (7) through (9), our results indicate that changes in the money rule do not have impacts on the real variables in the economy. And since this is true for the levels of these variables, it must hold for their variances as illustrated by (10). On the other hand, equations (7) through (9) show that with the exception of the real interest rate, the real variables in the economy do respond to unanticipated shocks. These are, of

[7] The solution to obtain (8) requires us to first obtain the forward solution to:

$$E_{t-1}(e_t - P_t) = -\frac{b_1}{b_2 - b_1}[E_{t-1}(e_{t+1} - P_{t+1}) + \bar{\imath}] - \frac{1}{b_2 - b_1}E_{t-1}y_t.$$

By repeated substitution, the above equation can be written as:

$$\begin{aligned} E_{t-1}(e_t - P_t) = &-\frac{1}{b_2 - b_1}\sum_{j=0}^{n}\left(-\frac{b_1}{b_2 - b_1}\right)^j (b_1\bar{\imath} - E_{t-1}y_{t+j}) \\ &+ \left(-\frac{b_1}{b_2 - b_1}\right)^{n+1} E_{t-1}(e_{t+n+1} - P_{t+n+1}). \end{aligned}$$

Thus, (8) requires the assumption that:

$$\lim_{n\to\infty} -\left(\frac{b_1}{b_2 - b_1}\right)^n E_{t-1}(e_{t+n} - P_{t+n}) \to 0.$$

The imposition of this terminal condition rules out the possibility of expected changes in future relative prices, which would lead individuals to demand infinite quantities of either foreign or domestic goods in the present period. This terminal condition is similar to that of Sargent and Wallace (1975, p. 248). Eq. (8) also requires that $E_{t-1}y_{t+j}$ not be too explosive, i.e., that a_2 not be too large.

course, the familiar results for aggregate models which incorporate rational expectations into a natural rate aggregate supply equation such as (1).[8]

The solutions (7) through (9) also have interesting implications for whether or not the domestic economy is independent of the rest of the world under a money rule. From (7) we find that the level of real output is independent of the world interest rate. So, from this perspective, our results agree with Parkin's (1977, p. 238) conclusion that the domestic economy is independent of the rest of the world under a money rule. However, the time paths of the real exchange rate and the real interest rate depend on the world interest rate but not on the money rule being followed. Thus, the domestic economy is not independent of the rest of the world in the sense that changes in the world interest rate will have real impacts on the domestic economy which cannot be offset through the money rule.[9]

The reason that a change in the world interest rate will have real impacts can best be seen by demonstrating that the opposite cannot occur. If changes in the world interest rate were to have no impacts on domestic real variables, an increase in $\bar{i}$ would have to be offset by an equal decrease in $E_{t-1}e_{t+1}$ in order to leave the real interest rate and the real exchange rate unchanged. However, the decrease in next period's exchange rate implies a decrease in next period's relative price of foreign and domestic goods and next period's demand for output. Consequently, current expectations of next period's price level will decrease. This reduces the demand for current period output since the real interest rate will be increased. However, since current output cannot change due to the form of (1), equilibrium must be restored through increases in the real exchange rate to offset the increase in the real interest rate.[10]

In order to illustrate the way in which unanticipated disturbances

[8] Explicit solutions for P_t and e_t could be obtained by solving the system of equations

$$b_2\begin{bmatrix} P_t \\ e_t \end{bmatrix} = \begin{bmatrix} c_3 & b_2 \\ -c_0 & b_2 \end{bmatrix} \begin{bmatrix} b_1[\bar{i} + E_{t-1}(e_{t+1} - e_t) - E_{t-1}(P_{t+1} - P_t)] + \nu_t - y_t \\ -c_2[\bar{i} + E_{t-1}(e_{t+1} - e_t)] + E_{t-1}m_t + \epsilon_t - \eta_t - c_1 y_t \end{bmatrix} \quad (11)$$

simultaneously using the Sargent and Wallace forward-solution technique. From (11) it is obvious that such solutions will include the $E_{t-1}m_{t+j}$, $j \geq 0$, as independent variables. Thus, the nominal variables in this economy will respond to changes in the money supply rule being followed.

[9] From (11) it is obvious that both $\bar{i}$ and policy parameters will appear in the solutions for the domestic price level, the exchange rate, and the nominal interest rate. Thus, the expected levels of these variables can be determined by the domestic monetary authorities regardless of the level of the world interest rate.

[10] Since we have assumed that the world price level is determinate and fixed, changes in $\bar{i}$ can only be caused by changes in the world real interest rate. Thus, it is not surprising to find the domestic real interest rate affected by such changes under flexible exchange rates.

affect the real and nominal variables under flexible exchange rates, we examine the effects of an unanticipated increase in the domestic nominal money supply. Differentiating (7) through (10) with respect to ϵ_t, we find that an unanticipated increase in the nominal money supply will increase real output, the exchange rate, and the price level while decreasing both the nominal and real interest rate. Further, if $a_1 > 0$, the increase in the price level will be less than the increase in the exchange rate so that the real exchange rate will increase.

These changes occur for the following reasons. The unexpected increase in the nominal money supply leaves expectations of current and future prices and exchange rates unchanged. However, it does cause money supply to be greater than money demand with a corresponding equal-valued excess demand for both domestic and foreign bonds which are perfect substitutes. The excess demand for bonds will put upward pressure on the exchange rate as agents attempt to sell domestic money for foreign money to buy foreign bonds. The rise in the exchange rate increases the demand for output through the IS curve. Thus, the effect of the unanticipated increase in the nominal money supply is to shift out the aggregate demand curve. There will then be an excess demand for output which will cause upward pressure on the domestic price level, and since the price level increase is unanticipated, it will cause the supply of output to increase unless $a_1 = 0$. Obviously, the rise in the price level must be less than the increase in the exchange rate otherwise there would be no shift in the aggregate demand curve.

Because of the recent discussion concerning whether the exchange rate will overshoot or undershoot its final level in response to a monetary disturbance,[11] we conclude the discussion of the money rule case by determining the time path of the exchange rate in response to an unanticipated change in the money supply. To facilitate this discussion, we assume a specific money rule of the simple autoregressive form

$$m_t = \rho m_{t-1} + \epsilon_t. \tag{12}$$

Under such a rule, the time path of the exchange rate, ignoring terms involving the world interest rate and errors other than ϵ_t, is

$$e_t = k(b_2 + a_1)\epsilon_t + \sum_{j=0}^{\infty} (\gamma_1 \rho^j + a_1 b_2 k \gamma_2 a_2^j)\epsilon_{t-1-j}, \tag{13}$$

where

[11] The overshooting phenomenon is obtained by Dornbusch (1976, 1978), although in the latter paper he finds that undershooting rather than overshooting occurs when price levels are allowed to adjust. The overshooting effect has also been found empirically by Frenkel (1978).

$$\gamma_1 = \frac{\rho}{1 - c_2(1 - \rho)}$$

and

$$\gamma_2 = a_2 \frac{c_0 - c_1[b_1(a_2 - 1) + b_2]}{[b_1(a_2 - 1) + b_2][1 + c_2(a_2 - 1)]}.$$

Differentiating (13) with respect to ϵ_t, we obtain the initial effect of the monetary shock to be

$$\frac{\partial e_t}{\partial \epsilon_t} = k(b_2 + a_1) > 0, \tag{14}$$

so that the immediate effect of the unanticipated increase in the money supply is to increase the exchange rate. The reasons for this effect are given in the above discussion of the impact effects of unanticipated changes in the money supply.

The effects of an unanticipated change in the money supply in period t on the exchange rate in subsequent time periods are obtained by differentiating (13) for period $t + j$ with respect to ϵ_t to obtain

$$\frac{\partial e_{t+j}}{\partial \epsilon_t} = \gamma_1 \rho^{j-1} + a_1 b_2 k \gamma_2 a_2^{j-1}, j \geq 1. \tag{15}$$

From (15) we first find that if both $|\rho|$ and $|a_2|$ are less than unity, unanticipated changes in the money supply will decay exponentially. We also find that unanticipated changes in the money supply in period t will not have impacts past the initial period unless there are feedback effects in either the money supply rule or the aggregate supply function, that is, unless either ρ or a_2 is nonzero.

However, when feedback effects are present in the money supply rule but not in the aggregate supply function ($a_2 = 0$ and $\rho > 0$), we find even though e_{t+1} will be above the long-run level of the exchange rate, we cannot determine whether it will be above or below e_t. Thus, in this case we find that even though the exchange rate will always initially overshoot its long-run level in response to an unanticipated money supply change, the overshooting actually may be larger in the period after the shock than in the period of the shock.

On the other hand, when we relax the assumption that $a_2 = 0$ and allow $0 \leq a_2 \leq 1$, we find that it is possible for the initial overshooting of the exchange rate in period t to be followed by undershooting in subsequent time periods. This result occurs because it is not possible to sign γ_2. Thus, our analysis indicates that even though an unanticipated money supply increase must initially increase the exchange rate, if lagged output terms appear in the aggregate supply function, the exchange rate may either undershoot or overshoot its long-run level in subsequent time periods.

Exchange Rate Rule

We now consider the case in which the monetary authority in the economy follows an exchange rate rule rather than a money supply rule. A monetary authority following an exchange rate rule sets a target value for the exchange rate ($\bar{e}_t$) according to some known rule. This rule could be of the linear feedback type. Alternatively, the rule could be that $\bar{e}_t = \bar{e}$ for all t, yielding the case of fixed exchange rates, or that the exchange rate appreciate or depreciate at a given rate over time. The important point for the analysis which follows, however, is not the particular form of the exchange rate rule. Rather it is that the rule used to determine the exchange rate be known and constant over time. In this case, the rational expectation of the exchange rate in period $t + 1$ will be $E_{t-1}(e_{t+j}) = \bar{e}_{t+j}$.

Of course, the monetary authority cannot establish $\bar{e}_t$ by fiat. Under an exchange rate rule the monetary authority must accomplish its objectives by buying or selling foreign or domestic reserves until $\bar{e}_t$ is achieved. However, since either action will affect the domestic money supply, under an exchange rate rule the domestic money supply becomes an endogenous variable. Further, given the structure of the economy, a given exchange rate rule will imply the time path of the domestic money supply and the monetary authority's holdings of international reserves.[12]

When m_t is endogenous, solving (1) through (4) for the price level yields

$$\begin{aligned} P_t = \{&b_1[\bar{\imath} + E_{t-1}(e_{t+1} - e_t) - E_{t-1}(P_{t+1} - P_t)] \\ &+ b_2 e_t + \nu_t + a_1 E_{t-1} P_t - a_2 y_{t-1} - \mu_t\}/(b_2 + a_1). \end{aligned} \tag{16}$$

Applying the E_{t-1} operator to both sides of (16) and subtracting the result from (16) yields

$$P_t - E_{t-1}P_t = (\nu_t - \mu_t + b_2\omega_t)/(b_2 + a_1), \tag{17}$$

where $\omega_t = e_t - E_{t-1}e_t$ is the innovation in the exchange rate in period t.[13] Substituting (17) into (1) yields the reduced-form solution for real

[12] Certain exchange rate rules might require the monetary authority to sell a quantity of international reserves exceeding their initial holdings. Such exchange rate rules would obviously not be feasible and are implicitly ruled out of the analysis. However, since the variance of real output under an exchange rate rule is independent of the actual rule chosen, the infeasibility of some exchange rate rules does not affect our results. I am indebted to a referee for this point.

[13] Since the monetary authority is intervening in the foreign exchange market, it is difficult to see how such innovations in the exchange rate could occur. Nonetheless, we have introduced ω_t into the analysis of the exchange rate rule case to make it symmetric to the money rule case in which the actual money supply could deviate from the desired money supply of the monetary authority by ϵ_t. However, since ω_t is a white-noise random error which is assumed to be uncorrelated with the other disturbances, the

output as

$$y_t = (b_2\mu_t + a_1\nu_t + a_1 b_2\omega_t)/(b_2 + a_1). \tag{18}$$

Using the Sargent-Wallace forward solution technique, the solution for the price level can be obtained from (16) as[14]

$$\begin{aligned} P_t = {} & \frac{b_1\bar{\imath}}{b_2} - \frac{a_2}{b_2 + b_1(a_2 - 1)} y_{t-1} + \frac{1}{(b_2 - b_1)} \sum_{j=0}^{\infty} \\ & \left(\frac{-b_1}{b_2 - b_1}\right)^j [b_1 E_{t-1} e_{t+j+1} + (b_2 - b_1) E_{t-1} e_{t+j}] \\ & - (\mu_t - \nu_t + b_2\omega_t)/(b_2 + a_1). \end{aligned} \tag{19}$$

The real exchange rate can be obtained by subtracting (19) from $e_t = E_{t-1}e_t + \omega_t$, and the real interest rate can be obtained by substituting for y_t and $e_t - P_t$ in (2).

The reduced-form solution for output (18) once again has the familiar neutrality result for models which incorporate rational expectations and the natural rate in the aggregate supply equation. However, since (19) depends on $E_{t-1}e_{t+j}$ terms, under an exchange rate rule both the real exchange rate and the real interest rate depend

---

inclusion of the ω_t does not modify the fact that e_t is exogenously determined in the exchange rate rule case.

[14] The solution to obtain (19) requires us to first obtain the forward solution to:

$$E_{t-1}P_t = -\frac{b_1}{b_2 - b_1}(E_{t-1}P_{t+1} + \bar{\imath} + E_{t-1}e_{t+1}) + E_{t-1}e_t - \frac{1}{b_2 - b_1}E_{t-1}y_t.$$

By repeated substitution, the above equation can be written as.

$$\begin{aligned} E_{t-1}P_t = {} & \sum_{j=0}^{n} \left(-\frac{b_1}{b_2 - b_1}\right)^j \left[\frac{b_1}{b_2 - b_1}(\bar{\imath} + E_{t-1}e_{t+j+1}) + E_{t-1}e_{t+j} \right. \\ & \left. - \frac{1}{b_2 - b_1}E_{t-1}y_{t+j}\right] + \left(-\frac{b_1}{b_2 - b_1}\right)^{n+1} E_{t-1}P_{t+n+1} \end{aligned}$$

Thus, (19) requires that.

$$\lim_{n\to\infty} -\left(\frac{b_1}{b_2 - b_1}\right)^n E_{t-1}P_{t+n} \to 0,$$

which rules out "speculative bubbles" in domestic goods Eq. (19) also requires that the process

$$\left(\frac{b_1}{b_2 - b_1}\right)E_{t-1}e_{t+j+1} + E_{t-1}e_{t+j} - \frac{1}{b_2 - b_1}E_{t-1}y_{t+j}$$

not be too explosive. These conditions are very similar to those required by Sargent and Wallace (1975). They differ from those necessary to obtain (8) because in the case of (8) we are solving for the real exchange rate and in the case of (19) we are solving for the domestic price level. It is also interesting to note that, whereas Sargent and Wallace have shown that in a closed economy the price level is indeterminate if the monetary authority follows an interest rate rule, we can obtain a determinate result if the monetary authority follows an exchange rate rule in an open economy.

upon the exchange rate rule being followed. The results that nominal variables (and the nominal money supply is included in this case) are affected by the policy parameters also continues to hold.

Further, from (18) we find that under an exchange rate rule the variance of real output is

$$\sigma^2_{y|E} = (b_2^2\sigma_\mu^2 + a_1^2\sigma_\nu^2 + a_1^2 b_2^2\sigma_\omega^2)/[(b_2 + a_1)^2(1 - a_2^2)], \qquad (20)$$

which once again is independent of the particular exchange rate rule being followed. However, we do find that the variance of output under an exchange rate rule differs from the variance of output under a money rule. One major difference is that under an exchange rate rule the variance of output is independent of any disturbances which occur in the money demand equation. The reason is the following. Under an exchange rate rule, the nominal interest rate is entirely determined by this rule and the world interest rate. With the interest rate thus fixed, real output and the price level are determined solely by the aggregate supply and IS curves. The only role of the money demand equation is to determine the level of the endogenous nominal money supply.

The solutions (18) and (19) also have interesting implications for whether or not the domestic economy is independent of the world interest rate under an exchange rate rule. Since the systematic part of the exchange rate rule (the $E_{t-1}e_{t+j}$ terms) appears in the solutions for both the real exchange rate and the real interest rate, the monetary authority can offset any changes in the world interest rate by changing its exchange rate rule. Specifically, if the monetary authority reduces the expected exchange rate in the next period by the amount of any increase in the world interest rate while holding the expected current exchange rate constant, the domestic economy will be independent of changes in the world interest rate. Such an exchange rate rule would keep both the domestic nominal interest rate and the current exchange rate constant. Consequently, the change in $\bar{i}$ would leave aggregate demand in the domestic economy unchanged with the result that no real variables would change.

Thus, we have found that whereas under a money rule the domestic economy will not be independent of the rest of the world, under a managed exchange rate system the domestic economy can be independent of the rest of the world in the sense that changes in the world interest rate will have no real effects. The difference in the two rules can be explained as follows. If changes in the world interest rate were to be neutralized under a money rule, the money supply would have to be systematically changed in response to changes in $\bar{i}$. However, since (2) through (4) are homogeneous of degree zero in the expected current money supply and the expected current and future price level

and exchange rate, any change in the expected money supply will be accompanied by proportional changes in the price level and exchange rate which leave the expected rate of change of the price level and the exchange rate unchanged. Consequently, under a money rule an increase in $\bar{i}$ will have real effects. Under an exchange rate rule, however, $E_{t-1}(e_{t+1} - e_t)$ can be modified without necessarily altering the current money supply, and an exchange rate rule can be devised to make the domestic economy independent of the world interest rate as discussed above.[15]

IV. Implications

Comparison of the variance of output under a money rule and under an exchange rate rule indicates that it is not possible to determine that one rule always dominates the other. Like the results in the traditional analysis of the merits of fixed versus flexible exchange rates, we see that which rule is preferred depends upon location and magnitude of the variance of the stochastic disturbances in the model. Nonetheless, we can determine some special cases in which one type of policy rule will yield a lower variance of output than the other.

We first consider the case in which the variance of the disturbance in the money demand function is large relative to the variances of the other disturbances. In this case, the variance of output will always be smaller if the monetary authority follows an exchange rate rule rather than a money rule, since under an exchange rate rule, real output is determined independently of the money market. Because fixed exchange rates are a special case of our exchange rate rule, this result is similar to Boyer's (1978) finding that a fixed exchange rate is optimal when shocks occur only in the money market and that is the market in which the authorities are intervening. Further, it corresponds to the results of Turnovsky (1976), Fischer (1977*a*), and Flood (1979) who find that money demand shocks do not affect output (consumption in the case of Fischer) under fixed exchange rates.

Next, we consider the case in which the variance of the disturbance to the aggregate supply curve is large relative to the variance of the other disturbances. Setting $\sigma_\nu^2 = \sigma_\eta^2 = \sigma_\epsilon^2 = \sigma_\omega^2 = 0$ in (11) and (20), we obtain

[15] On the other hand, if the monetary authority follows a fixed exchange rate rule, the domestic economy will be affected by changes in the world interest rate since changes in the world interest rate will not be offset by exchange rate changes. Thus, changes in $\bar{i}$ will have real effects since the change in the domestic nominal interest rate will shift the aggregate demand curve. Further, (19) shows that with fixed exchange rates, the price level will be determined solely by the level of the world interest rate and random shocks. Thus, a fixed exchange rate policy will yield no inflation on average unless there is an upward drift in the level of the world nominal interest rate.

$$\sigma^2_{y|M}/\sigma^2_{y|E} = (a_1 + b_2)^2/[b_2 + a_1(b_2c_1 + c_3)]^2. \tag{21}$$

Thus,

$$\sigma^2_{y|E} \left\{ \gtreqless \right\} \sigma^2_{y|M} \quad \text{as} \quad c_1b_2 + c_3 \left\{ \gtreqless \right\} 1. \tag{22}$$

This result is explained by the fact that the variance of the output response to an aggregate supply shock will be larger, the steeper is the slope of the aggregate demand curve. Under an exchange rate rule, the slope of the aggregate demand curve is

$$\left(\frac{\partial y_t}{\partial P_t}\right)_E = -b_2. \tag{23}$$

Under a money rule, the slope of the aggregate demand curve is

$$\left(\frac{\partial y_t}{\partial P_t}\right)_M = -b_2/(b_2c_1 + c_3). \tag{24}$$

Thus, when $c_1b_2 + c_3 < 1$, the slope of the aggregate demand curve will be steeper and the variance of output larger under a money rule. However, when $c_1b_2 + c_3 > 1$, the aggregate demand curve will have a steeper slope and the variance of output will be larger under an exchange rate rule.

These results do not agree with those of Fischer (1977*a*) and Lipschitz (1978) who find that fixed exchange rate rules are always preferred when the major sources of disturbances are shocks to output (supply) which are independent of demand. The differences arise because Fischer is concerned with the variance of consumption, which is defined to be the difference between nominal income and the balance of payments. He obtains his result since, under fixed rates, balance-of-payments outflows tend to offset part of the increase in consumption caused by the increase in income. In Lipschitz's model changes in the exchange rate cause wealth effects which tend to move in the same direction as the effects of the change in income. Thus, since fixed rates eliminate the wealth effects, they are variance minimizing.

The third case to consider is that in which the variance of the disturbance to the IS curve (caused, e.g., by large disturbances to the investment demand function) is large relative to the variance of the other disturbances. Setting $\sigma^2_\mu = \sigma^2_\eta = \sigma^2_\epsilon = \sigma^2_\omega = 0$, we obtain

$$\sigma^2_{y|M}/\sigma^2_{y|E} = c_3^2(a_1 + b_2)^2/[b_2 + a_1(b_2c_1 + c_3)]^2. \tag{25}$$

Under the presumption that c_3, the elasticity of money demand with respect to the exchange rate, is small, (25) indicates that a money rule will always yield a smaller variance of output against output demand disturbances. This result occurs because under an exchange rate rule

the nominal interest rate is fixed. Thus, any disturbance to the IS curve must be cleared by price-level and output changes. However, under a monetary rule part of the market equilibration can occur through changes in the interest rate. This lessens the amount of price-level change necessary to clear the market. Consequently, the amount of output response which must occur as a result of the shock to the IS curve will be smaller under a money rule. This is similar to Lipschitz's (1978) finding that if the major sources of disturbances are shocks to demand which are independent of real output, then an economy should use floating rates. Our result also agrees with Turnovsky's (1976) finding that, with a high degree of capital mobility, flexible exchange rates yield a higher degree of output stability than fixed exchange rates.

Finally, we consider the case in which the variances associated with the choice of the exchange rate regime itself are large relative to the variances of the other disturbances in the model. Setting $\sigma_\mu^2 = \sigma_\nu^2 = \sigma_\eta^2 = 0$ we obtain

$$\sigma_{y|M}^2/\sigma_{y|E}^2 = (\sigma_\epsilon^2/\sigma_\omega^2)\{(a_1 + b_2)^2/[b_2 + a_1(b_2c_1 + c_3)]^2\}. \tag{26}$$

Thus, (26) indicates that a money rule will be preferred to an exchange rate rule if unanticipated changes in the money supply cause smaller shifts in the aggregate demand curve than do unanticipated changes in the exchange rate and vice versa. This result can be obtained by solving the demand side of the model along the lines used to obtain the slopes of the aggregate demand curves for the two regimes.

V. Summary

We can summarize our findings in this paper as follows:

1. We find that we cannot determine that one type of monetary (or exchange rate) policy will always dominate the other. However, if we are able to determine that the variance of one type of shock to the economy is large relative to the other shocks, then the above analysis indicates that one type of rule can be shown to dominate the other.

2. On the other hand, there may be some merit in the discussion of the merits of freely fluctuating versus managed exchange rates based upon the degree of openness of the economy if the shocks to the domestic aggregate supply curve have the largest variance. The discussion in Tower and Willett (1976, pp. 44–45) indicates that b_2, the elasticity of the demand for output with respect to the relative price of domestic and foreign goods, is larger the more open the economy. Thus, the more open the economy, the more likely it is that a money

rule (freely fluctuating exchange rates) will yield a smaller variance of output in response to shocks to the aggregate supply curve.[16]

3. Since policy parameters do not appear in the variance of output under an exchange rate rule, fixed rates are as good as any other form of managed rate in terms of stabilizing output. Thus, arguments for managed rates as opposed to fixed rates must be based upon differences in price level variability or some other criteria. However, since the variance of output will be different under flexible than under managed exchange rates, output stability can be used as the criterion for deciding between the two types of exchange regimes.

4. Real variables in the economy can be insulated from changes in the world nominal interest rate under a managed exchange rate scheme, whereas the real exchange rate and the real interest rate will be affected by changes in the world interest rate under flexible exchange rates.

Appendix

We now solve the model for the symmetric information case in which information in period t is assumed to be available to agents in making their commodity and money demand decisions. Thus, we use the model

$$y_t = a_1(P_t - E_{t-1}P_t) + \mu_t \text{ (aggregate supply)} \tag{A1}$$

$$y_t = b_1[i_t - (E_tP_{t+1} - P_t)] + b_2(e_t - P_t) + \nu_t \text{ (IS)} \tag{A2}$$

$$m_t = c_0P_t + c_1y_t + c_2i_t + c_3e_t + \eta_t \text{ (LM)} \tag{A3}$$

$$i_t = \bar{i} + E_te_{t-1} - e_t \text{ (interest rate parity).} \tag{A4}$$

In (A1), we assume $a_2 = 0$ for simplicity.

We first consider the money rule case. When e_t is endogenous and letting $b_2^* = b_2 - b_1$ and $c_3^* = c_3 - c_2$, we solve (A1) through (A4) for P_t and obtain

$$\begin{aligned} P_t = k^*\{&(a_1c_3 + b_2^*c_1)E_{t-1}P_t + c_3^*b_1[\bar{i} + E_t(e_{t+1} - P_{t+1})] \\ &+ b_2^*m_t - b_2^*c_2(\bar{i} + E_te_{t+1}) + \theta_t^*\}, \end{aligned} \tag{A5}$$

where $\theta_t^* = [-(c_3^* + b_2^*c_1)\mu_t + c_3^*\nu_t - b_2^*\eta_t]$ and $k_t^* = [b_2^*(1 - c_2) + a_1c_3^* + c_1a_1b_2^*]^{-1}$.

Since (A1) must hold for all periods, it must hold for period $t + j$, where $j > 0$. Thus,

$$y_{t+j} = a_1(P_{t+j} - E_{t+j-1}P_{t+j}) + \mu_{t+j}. \tag{A1'}$$

Applying the operator E_t to both sides of (A1′) yields

$$E_ty_{t+j} = 0, \qquad j > 0. \tag{A6}$$

[16] Since Flood (1979) assumes that purchasing power parity holds, he is concerned with the case in which $b_2 \to \infty$. He finds that var $(P_t - E_{t-1}P_t)$ is not affected by changes in σ_μ^2 under fixed exchange rates, whereas it is affected by changes in σ_μ^2 under flexible exchange rates. Taking the variances of (6) and (17) and letting $b_2 \to \infty$ shows that Flood's result also holds for our model.

Equation (A2) must also hold for all periods. Applying the E_t operator to both sides of (A2) for period $t + j$ and using (A4) and (A6) yields

$$0 = b_1[\bar{\imath} + E_t(e_{t+j+1} - P_{t+j+1})] + b_2^* E_t(e_{t+j} - P_{t+j}). \tag{A2'}$$

This first-order difference equation in $E_t(e_{t+j} - P_{t+j})$ can be solved forward, yielding

$$E_t(e_{t+j} - P_{t+j}) = -b_1\bar{\imath}/b_2^*, \qquad j > 0. \tag{A7}$$

Equation (A3) must also hold for all periods. Applying the E_t operator to both sides of (A3) for period $t + j$ and using (A4) and (A6) yields

$$\begin{aligned} E_t m_{t+j} = &-c_0 E_t(e_{t+j} - P_{t+j}) + c_2\bar{\imath} + c_2 E_t e_{t+j+1} \\ &+ (1 - c_2)E_t e_{t+j+1}. \end{aligned} \tag{A3'}$$

Equation (A3′) can be solved forward, and using (A7) we obtain

$$E_t e_{t+j} = -(c_2 + c_0 b_1/b_2)\bar{\imath} + (1 - c_2)^{-1} \sum_{i=0}^{\infty} \left(\frac{-c_2}{1 - c_2}\right)^i E_t m_{t+j+i}. \tag{A8}$$

After substituting for $E_t(e_{t+1} - P_{t+1})$ using (A7) and $E_t(e_{t+1})$ using (A8), we apply the E_{t-1} operator to both sides of (A5) and subtract the result from (A5) to obtain

$$P_t - E_{t-1}P_t = k^*(\theta_t^* + b_2^*\epsilon_t). \tag{A9}$$

Substituting (A9) into (A1) yields the solution for real output as

$$y_t = k^*[b_2^*(1 - c_2)\mu_t + c_3^*\nu_t + b_2^*(\epsilon_t - \eta_t)]. \tag{A10}$$

And from (A10) we find that the variance of output under a money rule will be

$$\sigma^2_{y|M} = (k^*)^2[(b_2^*)^2(1 - c_2)^2\sigma_\mu^2 + (c_3^*)^2\sigma_\nu^2 + (b_2^*)^2(\sigma_\epsilon^2 + \sigma_\eta^2)]. \tag{A11}$$

Further, if we were to solve for the real exchange rate and the real interest, we would once again find that they were independent of the systematic part of monetary policy but not independent of the world interest rate. Thus, these results do not differ much from those in the text.

We next consider the case of the exchange rate rule. When m_t is endogenous, we solve (A1) through (A4) for P_t and obtain

$$\begin{aligned} P_t = \{&a_1 E_{t-1}P_t - \mu_t + b_1[\bar{\imath} + E_t(e_{t+1} - P_{t+1})] \\ &+ \nu_t + b_2^* e_t\}/(a_1 + b_2^*). \end{aligned} \tag{A12}$$

Following the same procedures used to obtain (A9), we find that under an exchange rate rule

$$P_t - E_{t+1}P_t = (-\mu_t + \nu_t + b_2^*\omega_t)/(b_2^* + a_1). \tag{A13}$$

Substituting (A13) into (A1) yields

$$y_t = (b_2^*\mu_t + a_1\nu_t + a_1 b_2^*\omega_t)/(b_2^* + a_1). \tag{A14}$$

Therefore, the variance of output under an exchange rate rule is

$$\sigma^2_{y|E} = [(b_2^*)^2\sigma_\mu^2 + a_1^2\sigma_\nu^2 + (a_1 b_2^*)^2\sigma_\omega^2]/(b_2^* + a_1)^2. \tag{A15}$$

If we were to solve for the real exchange rate and the real interest rate, we would find they have the same properties with respect to systematic changes

in the exchange rate rule and changes in the world interest rate as do the solutions in the text.

Further, when expectations are based upon information in period t, the slopes of the aggregate demand curve under a money rule and an exchange rate rule are

$$\left(\frac{\partial y_t}{\partial P_t}\right)_M = -b_2^*(1 - c_2)/(c_3^* + c_1 b_2^*) \tag{A16}$$

and

$$\left(\frac{\partial y_t}{\partial P_t}\right)_E = -b_2^*, \tag{A17}$$

respectively. Thus, the cases in which one type of rule will dominate the other are unchanged from the text.

The only result which changes due to the change of timing of expectations is the time path of the response of the exchange rate to an unanticipated increase in the money supply. And letting the aggregate supply curve be (1) and the money rule be (12), it can be shown that the only difference in the effects of unanticipated changes in the money supply on the exchange rate due to the change in the timing of expectations will be in the $\partial e_t/\partial \epsilon_t$ term.

References

Barro, Robert J. "Rational Expectations and the Role of Monetary Policy." *J. Monetary Econ.* 2 (January 1976): 1–32.

Bilson, John F. O. "Rational Expectations and the Exchange Rate." In *The Economics of Exchange Rates*, edited by Jacob A. Frenkel and Harry G. Johnson. Reading, Mass.: Addison-Wesley, 1978.

Boyer, Russell S. "Optimal Foreign Exchange Market Intervention." *J.P.E.* 86, no. 6 (December 1978): 1045–55.

Dornbusch, Rudiger. "Expectations and Exchange Rate Dynamics." *J.P.E.* 84, no. 6 (December 1976): 1161–76.

———. "The Theory of Flexible Exchange Rate Regimes and Macroeconomic Policy." In *The Economics of Exchange Rates*, edited by Jacob A. Frenkel and Harry G. Johnson. Reading, Mass.: Addison-Wesley, 1978.

Fischer, Stanley. "Stability and Exchange Rate Systems in a Monetarist Model of the Balance of Payments." In *The Political Economy of Monetary Reform*, edited by Robert Z. Aliber. Montclair, N.J.: Allanheld, Osmun, 1977. (*a*)

———. "Wage Indexation and Macroeconomic Stability." In *Stabilization of the Domestic and International Economy*, edited by Karl Brunner and Allan H. Meltzer. Amsterdam: North-Holland, 1977. (*b*)

Flood, Robert P. "Capital Mobility and the Choice of Exchange Rate System." *Internat. Econ. Rev.* 20 (June 1979): 405–16.

Frenkel, Jacob A. "A Monetary Approach to the Exchange Rate: Doctrinal Aspects and Empirical Evidence." In *The Economics of Exchange Rates*, edited by Jacob A. Frenkel and Harry G. Johnson. Reading, Mass.: Addison-Wesley, 1978.

Gordon, Robert J. *Macroeconomics*. Boston: Little, Brown, 1978.

Gray, Jo Anna. "Wage Indexation: A Macroeconomic Approach." *J. Monetary Econ.* 2 (April 1976): 221–35.

Leiderman, Leonardo. "Expectations and Ouput-Inflation Tradeoffs in a Fixed-Exchange-Rate Economy." *J.P.E.* 86, no. 6 (December 1979): 1285–1306.

Lipschitz, Leslie. "Exchange Rate Policies for Developing Countries: Some Simple Arguments for Intervention." *Internat. Monetary Fund Staff Papers* 25 (December 1978): 650–75.

Lucas, Robert E., Jr. "Some International Evidence on Output-Inflation Tradeoffs." *A.E.R.* 63 (June 1973): 326–34.

McCallum, Bennett T. "Dating, Discounting, and the Robustness of the Lucas-Sargent Proposition." *J. Monetary Econ.* 4 (January 1978): 121–29.

Parkin, Michael. "The Transition from Fixed Exchange Rates to Money Supply Targets." *J. Money, Credit and Banking* 9, no. 1, pt. 2 (February 1977): 228–42.

Sargent, Thomas J., and Wallace, Neil. "'Rational' Expectations, the Optimal Monetary Instrument, and the Optimal Money Supply Rule." *J.P.E.* 83, no. 2 (April 1975): 241–54.

Tower, Edward, and Willett, Thomas D. "The Theory of Optimum Currency Areas and Exchange Rate Flexibility." Special Papers in International Economics, no. 11, Princeton Univ., Dept. Econ., 1976.

Turnovsky, Stephen J. "The Relative Stability of Alternative Exchange Rate Systems in the Presence of Random Disturbances." *J. Money, Credit and Banking* 8 (February 1976): 29–50.

Point Spreads versus Odds

Gilbert W. Bassett, Jr.

University of Illinois at Chicago Circle

Why do bookies use the point-spread device for wagering on football games when odds betting on the winner is also feasible? One possibility is that point-spread betting is more fun than odds betting. The existence of point spreads, however, does not logically entail this "more fun" explanation. Given simple specifications of beliefs and betting behavior, and given that point spreads and odds are equally thrilling, it is demonstrated that the point-spread game can exist as the game produced by a profit-maximizing bookie. The disparate beliefs which create the demand for some form of betting market can also explain the particular betting game produced by the bookie.

I

The latest point spread on the Bears-Rams football game this weekend lists the Bears as nine-point underdogs. For a 10-cent fee a bookie will accept an even-odds one-dollar wager on the event $B + 9 > R$, where B and R are the Bears' and Rams' final point totals. Based on your knowledge of the teams' relative strengths, you assign a (personalistic) probability of .7 to the event $B + 9 > R$ (and a probability of zero to $B + 9 = R$). Hence, a one-dollar (plus the 10-cent bookie fee) wager on the Bears has an expected monetary value (EMV) of $.7(1 - .1) - .3(1 + .1)$, or a positive 30 cents. Given the

An earlier version of this paper was presented at the Econometric Society Meeting, New York City, December 1977. I want to acknowledge helpful comments from a referee at this *Journal*.

[*Journal of Political Economy*, 1981, vol. 89, no. 4]
© 1981 by The University of Chicago 0022-3808/81/8904-0001$01.50

alternative of not betting, you as an *EMV* maximizer will wager that the Bears will beat the spread.

Point-spread schemes characterize virtually all of the wagering in football betting markets. So-called odds betting, however, is also possible and can provide for any particular individual an *EMV* identical to the point-spread game. If the odds are 6 : 1 on a Bear victory ($B > R$) and if you assign a probability of .2 to this event (and a probability of zero to $B = R$), then the *EMV* of a one-dollar (plus the 10-cent bookie fee) Bears bet is $.2(6 - .1) - .8(1 + .1)$, which is again 30 cents. Ceteris paribus, you are indifferent between the point-spread and odds wagers. Hence the question, Why do bookies produce the point-spread betting game?

The existence of point spreads is commonly attributed to bettor preferences. This explanation is implicit in news articles on point spreads, and it is also the most frequently offered explanation in the very informal settings where I have posed the point spreads versus odds question. Under this view bettor preferences are thought to depend on more than *EMV* or expected utility—utility being defined on income or wealth. Point spreads are said to be more fun or thrilling than odds so that bettors are willing to forgo expected return with odds in order to play the point-spread game. This appeal of point spreads could reflect a general preference for even-odds wagers, or it could be simply that point spreads are more complementary with viewing a football game. But the bias in favor of even-odds wagers does not extend to other lotteries, and many football bettors presumably wager on games they do not watch. It would appear that the majority (in the informal settings) think the intrinsic delight in beating the spread is obvious and requires no deeper analysis. The answer to the point spreads versus odds question, then, is simply that point spreads are more fun than the odds betting game.

Bettor preferences provide an answer to the point spreads versus odds question, but the explanation posits rather curious tastes which cannot be understood in terms of an *EMV* or expected utility decision criterion. It is possible, however, that preferences are in fact consistent with some more traditional decision criterion, and the stated views merely reflect the difficulty of imagining football betting without the point-spread device. Alternatively, the stated views might derive from the belief that bettor preferences are the only logical explanation consistent with the existence of point spreads. Without denying the possible empirical validity of the bettor-preferences explanation, the intent here is to generate alternative hypotheses for point spreads in a more traditional setting where bettors do not consider point spreads more fun. Disparate beliefs regarding the random variable $X = R - B$ are one reason for the existence of some

betting market, and they are examined here as also providing an explanation for the particular betting game produced by the bookie.

The market of potential bettors will be assumed fixed, and each bettor will be described by a probability distribution representing beliefs regarding X. In deciding on wagers the bettors are assumed to use an *EMV* criterion. The thrill of betting is therefore supposed constant over alternative betting games, and risk preference or aversion is ignored. (Bettors might be imagined to risk only a small fraction of their income or wealth.) Betting is facilitated by a profit-maximizing bookie who is paid a fee by participating bettors in exchange for either a point-spread, an odds, or a combined-betting game. To insulate himself from the outcome of X the bookie will also establish equilibrium parameters for the betting games. (The betting market is of some independent interest because it contains an agent, the bookie, who has an *incentive* to establish these equilibrium values.) The betting market is therefore envisioned as arising for purely speculative reasons. The market exists for the sole purpose of expediting the exchange of wagers whose *EMV*s are positive rather than because of variations in initial allocations of physical-location-dated commodities or the possibility of utility-increasing risk transfers. Wagering is rational in that each bettor is made better off according to his *EMV* criterion even though it is certain that the aggregate of bettors will lose money to the bookie.

The bookie's profit-maximizing objective translates in the games described below into the objective of maximizing the betting pool—the total dollars wagered on the football teams. Given the assumptions that no wagering scheme is intrinsically more fun than any other and that all individuals wager rationally according to an *EMV* criterion, and given the example above which illustrates that for any individual there is some odds game in which betting behavior is identical to the point-spread game, it might seem that the bookie would be indifferent as to which game to produce. In the model, point spreads would then exist merely as the outcome of the bookie's random choice from a set of equivalent betting games. If a deterministic rationale for point spreads is sought, the model would then have to be abandoned in favor of the bettor-preference explanation. We will see, however, that depending on beliefs, certain betting games generate greater profits for the bookie so that an appeal to the supposed delights of point spreads is not logically necessary in order to understand the bookie's production decision.

The rationale for assuming that individual beliefs are consistent with a probability distribution comes from the personalistic or subjective theory of probability, and Winkler (1971) has found some empirical support for the theory in the football context. In the economics

literature, I know of only two references dealing directly with football betting.[1] Pankoff (1968) used football betting to analyze issues relating to the efficient-market hypothesis. Smith (1971) examined betting markets because of the equilibrating role played by the bookie and because football wagering could provide an unusual framework for examining decision making under uncertainty. He used the expected-utility hypothesis to analyze how betting behavior under various betting schemes might be influenced by wealth, beliefs, and the pleasures of gambling. While Smith did consider behavior under point spreads and odds games and his model is similar to the one presented here, he did not examine the conditions under which one betting game is produced over another.

No attempt is made here to associate differing opinions with different or incomplete information among potential bettors. In fact, potential bettors might be envisioned as having virtually identical information on the football teams' current situation and past history. Still, bettors are not presumed to agree on the distribution of X. While my casual observations suggest that this is a reasonable working hypothesis, it does conflict with the spirit of rational expectations where differing opinions are associated with and analyzed in terms of some more fundamental differences among individuals. Besides casual empiricism, the fact that the probability assigned to a future event like the winner of a football game cannot be as easily inferred as the probability assignments in, say, roulette might also serve to justify the existence and persistence of differing opinions. Despite being able to continuously update their theories by observing the outcomes of football games, potential bettors do not reach agreement. Theories are sufficiently vague or depend on so many exogenous factors that, while each bettor possesses a well-defined distribution for X, difference of opinion exists and persists. Difference of opinion is what makes for a horse race and is what is examined here as an explanation for the point-spreads versus odds question.

The plan of the rest of the paper is as follows. Section II describes the bookie's objectives and production possibilities, that is, the games which can be offered to bettors. A simple parameterization of market beliefs and a betting criterion are presented in Section III. The assumptions are adopted in order to illustrate how point spreads might beat odds and also to determine in a simple model the conditions for the optimality of a betting game. Section IV begins the exercise by working through a special case. Propositions in Section V express equilibrium values for the betting games and bookie profits as a function of the parameters describing market beliefs. Section VI

[1] I am indebted to Stephen Salant for making me aware of these references.

then presents conditions under which point spreads beat odds. The assumptions on beliefs and betting criteria are relaxed in Section VII and another example is presented. The final section contains some concluding remarks.

II

The bookie brings together individuals with disparate beliefs regarding X by producing a betting game called spread-odds (s,z). Depending on the bookie's choice of the spread parameter s and the odds parameter z, the spread-odds (s,z) game is either the point-spread game, the odds game, or a game combining features of both point spreads and odds. The spread-odds (s,z) device, apparently first used by Smith (1971), is convenient for stating results and for obtaining conditions under which point spreads beat odds and any other combined-betting game.

The bookie collects a fee or commission c for each bet purchased in the spread-odds (s,z) game. The commission is assumed fixed so that s and z are the bookie's only choice variables.[2] Furthermore, the bookie's administrative and other costs per bet are assumed to not depend on the particular spread-odds (s,z) game which is produced.

Bettors are offered a Rams bet or a Bears bet in the spread-odds (s,z) game (bettors will also have the option of not betting). The bets are a function of (s,z) where $-\infty < s < \infty$ and $0 < z < 1$. Each bet is described by its net gain, its net loss, and the event under which the gain is realized. The spread-odds (s,z) bets are: Rams bet, $[2z - c : 2(1 - z) + c, X > s]$; Bears bet, $[2(1 - z) - c : 2z + c, X < s]$. An individual who purchases one Rams bet will pay the bookie $2(1 - z) + c$ dollars and win two dollars gross, $2z - c$ dollars net, if $X > s$. If $X < s$ the Rams bettor loses his initial payment of $2(1 - z) + c$ dollars.[3] Similarly, a Bears bettor wins $2(1 - z) - c$ dollars per bet if $X < s$ and loses $2z + c$ dollars per bet if $X > s$.[4]

[2] In practice the bookie's commission is also a choice variable whose equilibrium value is determined by the betting industry's market structure. For an interesting study in a slightly different context of how the choice of c can influence (a race track's) profit, see Gruen (1976). To partially justify the c fixed assumption used here, one might suppose a competitive betting industry in long-run equilibrium with the average cost per wager equal to c.

[3] If $X = s$ all bets are off. This differs from some betting games in which the bookie wins all ties.

[4] The odds parameter z is a transformation of what is usually referred to as "the odds." E.g., if $\bar{z} = 2/3$ and $c = 1/99$, then the Rams bet is established at (131/99:67/99) and the Bears bet at (65/99:133/99). In the usual parlance this represents approximately 1.96 to 1 odds and 1 to 2.05 odds, respectively, the slight asymmetry in the odds being the result of the fee c. In order that the net gains and losses in the spread-odds (s,z) game be positive, the domain of z must be slightly restricted to the interval $(c/2, 1 - c/2)$.

The bookie's choice of a spread-odds (s,z) game is restricted to the set of (s,z) which makes profit independent of the outcome of the football game. This domain for (s,z) is called the equilibrium set and is given by $V = [(s,z) \mid \Lambda_R(s,z:c) = \Lambda_B(s,z:c)]$, where Λ_R and Λ_B denote the number of Rams and Bears bets purchased at the spread s, the odds z, given the commission c.

To see that (s,z) in V makes profit independent of X, consider the betting pool (the total dollars wagered on the Rams and Bears), $[2(1 - z) + c]\Lambda_R + [2z + c]\Lambda_B$. For (s,z) in V the pool is $(1 + c)\Lambda$ where $\Lambda = \Lambda_R + \Lambda_B$. From the pool the bookie returns $2\Lambda_R$ dollars to Rams bettors if $X > s$, and at (s,z) in V this is identical to the $2\Lambda_B$ dollars returned to Bears bettors if $X < s$. The bookie's profit is therefore $c\Lambda(s,z:c)$ with certainty for each (s,z) in V.

A point-spread game is spread-odds $(s, ½)$; that is, z is fixed at ½. With this restriction, wagers are made at (neglecting c) even odds on $X > s$ or $X < s$. The equilibrium point spread is $\bar{s}$ where $(\bar{s}, ½)$ is in V, spread-odds $(\bar{s}, ½)$ is the point-spread game which the bookie may produce, and the profit from producing this game is $c\Lambda(\bar{s}, ½:c)$.

An odds-betting game is spread-odds $(0, z)$ so that bets are made on the events $X > 0$ or $X < 0$, Rams win or Bears win. The equilibrium odds is denoted by $\bar{z}$ where $(0,\bar{z})$ is in V, spread-odds $(0,\bar{z})$ is the odds game which the bookie might produce, and the profit from this game is $c\Lambda(0,\bar{z}:c)$.[5]

The profit-maximizing bookie who is permitted to produce any spread-odds (s,z) game will select an (s^*,z^*) in V so that $\Lambda(s^*,z^*:c)$ is a maximum, and spread-odds (s^*,z^*) is called an optimal betting game. If Λ is constant over all (s,z) in V then every betting game is equivalent and optimal, and the bookie's profit is identical no matter which game is produced. This situation corresponds to the initial hypothesis under which the bookie is indifferent between alternative betting games. If $\Lambda(s^*,z^*:c) > \Lambda(s,z:c)$ for some (s,z) in V, then some betting games will not be produced because they result in lower profits than an optimal game. Finally, when Λ reaches a maximum for a unique (s,z) in V, there is a unique optimal game, and profit maximizing by the bookie completely determines the spread-odds (s,z) game which is produced.

The bookie who is constrained to either point spreads or odds will produce the former when $\Lambda(\bar{s}, ½:c) > \Lambda(0,\bar{z}:c)$. When this holds we will say that point spreads beat odds. Notice that point spreads, even if it beats odds, need not be an optimal game. Finally, an extreme case to be considered in Section VII occurs when point spreads beat odds

[5] In order that the point-spread and odds games are not identical, it is supposed that $\bar{s}$ is not equal to zero.

and $\Lambda(0,\bar{z}:c) = 0$; that is, no one purchases the odds wager. In this case we will say that point spreads versus odds is no contest.

Remarks

1. In practice a bookie who fails to select an (s,z) in V will try to exchange some of his wagers with another bookie or will begin to alter the spread and odds parameters. Hence, all bets may not be purchased at the same (s,z). In the model here, however, all bets are assumed to be made at an (s,z) in V. One might suppose that the bookie's skill is his ability to read the market and know the (s,z) such that $\Lambda_R = \Lambda_B$.

2. The equilibrium set corresponds to one possible measure of consensus regarding the distribution of X. If $\pi(s) = z$ for (s,z) in V, then π is a market probability distribution in the sense that the fraction of the betting pool (excluding the bookie fee) wagered on the event $X < s$ is z, and the fraction wagered on $X > s$ is $1 - z$. Bookies who fail to select (s,z) in V become bettors themselves and effectively are betting that their opinion is more accurate than this market consensus. Natural selection probably eliminates these bookies quickly either because they have become wealthy and retire or, more likely, because they have become financially or physically bankrupt, the latter bankruptcy being a likely consequence of not paying winning bettors.

3. In the standard point-spread game the bookie's commission (called the vigorish) results in spread bets being offered at 10:11 odds. With the spread-odds (s,z) notation this means that the point-spread bet is offered at odds of 20/21:22/21 and $c = 1/21$.

III

The set of potential bettors is denoted by N and remains fixed throughout. Associated with each individual i in N is a probability distribution $F_i(x)$ and a betting function; $F_i(x)$ is the ith bettor's probability assignment to the event $X < x$, and the betting function along with $F_i(x)$ determines the number of Rams or Bears bets purchased by bettor i.

Aside from the example in Section VII, beliefs will be assumed identical except for a location parameter. With this simple specification $F_i(x) = G(x - \mu_i)$, and G is further taken to be a symmetric (about zero, $G(x) = 1 - G(-x)$) distribution with continuous derivative $g > 0$. Given G, the ith bettor's beliefs are completely determined by μ_i, the bettor's opinion regarding the median-mathematical expectation of X. The distribution of the μ_i over the set of potential bettors is approximated with N large by the continuous and increasing dis-

tribution $M(y - \mu_M)$—the proportion of potential bettors who believe the median-mathematical expectation of X is less than y. The distribution M is further assumed symmetric about zero so that μ_M is the median or average of all bettors' beliefs regarding the median-mathematical expectation of X. Market beliefs under this simple representation are therefore a location family of distributions depending on G, M, and μ_M.

A potential bettor will participate in the betting market by purchasing a positive quantity of the Rams or Bears bet if and only if the *EMV* of the corresponding bet is positive. With this requirement the set of potential bettors may be partitioned as follows.

Given a spread-odds (s,z) game the *EMV* of one Rams bet for bettor i is $2z - c - 2G(s - \mu_i)$. The requirement that this be positive can be expressed directly in terms of μ_i as $\mu_i > s - Q(z - c/2)$, where $Q(p)$ is the (increasing) inverse or pth quantile of G. Making similar calculations for the Bears bet gives the partitioning, Rams bettors = $[i \text{ in } N \mid \mu_i > s - Q(z - c/2)]$, Bears bettors = $[i \text{ in } N \mid \mu_i < s - Q(z + c/2)]$; and the nonbettors are those individuals whose μ_i is the interval $[s - Q(z + c/2), s - Q(z - c/2)]$. This holds for arbitrary, not necessarily equilibrium, values for s and z. To determine V requires further assumptions which describe the quantity of wagers purchased when the *EMV* is positive.

Let λ_i^R and λ_i^B denote the number of Rams and Bears bets purchased by the ith individual. The betting rule assumed here is $\lambda_i^R = \lambda[\mu_i - s + Q(z - c/2)]$, and $\lambda_i^B = \lambda[s - \mu_i - Q(z + c/2)]$, where $\lambda(w)$ is a positive and nondecreasing function for $w > 0$ and is zero otherwise. This rule is consistent with the positive *EMV* requirement. If $\lambda(w) = 1$ for $w > 0$, bettors are restricted to a single bet when its *EMV* is positive. If $\lambda(w)$ is increasing for $w > 0$ the quantity of bets purchased is based not on *EMV* directly but on the difference between μ_i and $s - Q(z - c/2)$ or $s - Q(z + c/2)$, depending on which wager's *EMV* is positive. If $\mu_1 > \mu_2 > s - Q(z - c/2)$ the *EMV* of one Rams bet is greater for bettor 1 and this bettor purchases more Rams bets than bettor 2, $\lambda_1^R > \lambda_2^R > 0$. With this betting rule we will have $\lambda_i^R = \lambda_j^B > 0$ (i and j have equal but opposite opinions) when μ_i and μ_j are equidistant but on opposite sides of the interval which describes the nonbettors.

Remarks

1. The assumption of a large number of individuals with well-defined distributions defined on $X = R - B$ is an empirical proposition which is probably not unrealistic (see Winkler 1971). If X represents an arbitrary future event, however, the assumption need not be

satisfied, and the model will not be applicable.[6] For example, if individuals do not possess a subjective distribution on the electoral (or popular) vote point spread, then the model cannot explain why odds beats point spreads in wagering on the outcome of a presidential election.

2. Under the football game interpretation, each potential bettor knows that $X = R - B$ is an integer-valued random variable. Therefore, there is the assumption that the outcomes of $R - B$ are sufficiently numerous that $F_i(x)$ may be approximated by a continuous distribution. In other cases this may be far from adequate, for example, baseball and soccer games. In the limit where the winner scores one and the loser zero, no point-spreads versus odds question will arise.

3. The location family of distributions is a very restrictive assumption which undoubtedly does not allow for the full diversity of actual beliefs. The assumption is adopted because it permits easy calculation of optimal game conditions and thereby will illustrate, albeit in a restrictive setting, how one betting game can be favored over another.

4. The assumed betting criterion is also adopted because it is convenient. With a location family of market beliefs, neither the equilibrium set nor the optimal game conditions will be found to depend on λ. But the criterion might also be fairly realistic since bettors often rationalize their wagers by appealing to the difference between the quoted spread and the spread at which they would decide not to wager.

IV

Before examining the conditions for an optimal game given an arbitrary location family of market beliefs, it is instructive to consider the following special case.

Suppose G is a uniform distribution on the interval $(-10, 10)$; $G(x) = .5 + .05x$. Let $c = .1$, $\mu_M = 3$, and let M be an arbitrary distribution satisfying the previous assumptions.

A Rams bettor is anyone who finds the *EMV* of one Rams bet to be positive or, on making the calculations, $\mu_i > s - 20z + 11$. Under the assumed betting criterion such a bettor purchases $\lambda_i^R = \lambda(\mu_i - T - 4)$ Rams bets, where we let T denote $s - 20z + 7$. Summing (integrating) over all such purchases gives

[6] The assumption can fail in either of two ways. Beliefs might be so vague that individuals feel completely incapable of assigning probabilities to future events. Alternatively, beliefs might be inconsistent with a probability distribution, e.g., if one assigned probabilities of .3 and .8 to the respective events $X < 0$ and $X > 1$.

$$\Lambda^R = n\int_{\mu>T+4} \lambda(\mu - T - 4)dM(\mu - 3)$$

$$= n\int_{w>T+1} \lambda(w - T - 1)dM(w),$$

where n is the number of potential bettors. Making similar calculations gives

$$\Lambda^B = n\int_{w<T-1} \lambda(-w + T - 1)dM(w).$$

This expresses the total quantity of Rams and Bears bets purchased given an arbitrary spread-odds (s,z) game.

To solve for the (s,z) in V use the symmetry of M and obtain

$$\Lambda^R - \Lambda^B = n\int_{w>T+1} \lambda(w - T - 1)dM(w)$$

$$- n\int_{w>-T+1} \lambda(w + T - 1)dM(w).$$

At $T = 0$ this is zero and, since M is increasing, $\Lambda^R - \Lambda^B = 0$ only if $T = 0$. Hence, $T = 0$ or $s(z) = 20z - 7$ gives the equilibrium set; at any $[s(z), z]$ the bookie's profit will not depend on the realization of X. The equilibrium point spread would be $s(\frac{1}{2}) = 3$, and the equilibrium odds is a solution to $s(z) = 0$ or $z = .35$.

Substituting $s(z)$ for s in the expressions for Λ^R and Λ^B gives the representation of profits for (s,z) in V as

$$c(\Lambda_R + \Lambda_B) = 2cn\int_{w>1} \lambda(w - 1)dM(w),$$

a term which does not depend on (s,z). So why does the bookie produce one game, say point spreads, over another? No (deterministic) answer is to be found in this example. The bookie's profit is identical in all spread-odds (s,z) games, every game is optimal, and disparate beliefs as represented here do not offer an explanation for the production of a particular betting game. In a world where the assumptions of this example are satisfied, the existence of point-spread betting would merely represent the bookie's random choice from the set of all spread-odds (s,z) games.

To put this result differently, consider the point-spread game with $\bar{s} = 3$. Instead of computing the equilibrium odds, mimic the example in the introduction and ask each bettor to compute the odds game at which his behavior would be identical to behavior under point spreads. For example, a Rams bettor with point spreads will purchase $\lambda(v_i)$ bets where $\mu_i - 4 = v_i > 0$. An odds game in which the same

number of Rams bets would be purchased is found by solving $\mu_i - 11 - 20z = v_i$, or $z = .35$. Every Rams and, with similar calculations, every Bears bettor will obtain this same solution, and at $z = .35$ every nonbettor under point spreads remains a nonbettor under odds. The behavior of each individual is therefore invariant under the transformation from $(s,z) = (3, \frac{1}{2})$ to $(0, .35)$. Since $(3, \frac{1}{2})$ is in V and because no behavior is altered at $(0, .35)$, it follows that .35 is the equilibrium odds. The point-spread and odds games are therefore equivalent.

In this example any individual can find an equalizing odds game, and this is generally true as long as bettors use an *EMV*-based decision criterion. What makes this example special and misleading is that every individual arrives at the same equalizing odds of $z = .35$. When G is not uniform we will see that not all bettors compute the same equalizing odds, and the invariance of betting behavior as (s,z) varies over V will no longer hold.

V

For the spread-odds (s,z) game with arbitrary (s,z), the *EMV*-based betting criterion provides a straightforward characterization of Λ^R and Λ^B as a function of G, M, and μ_M. The equilibrium set can then be obtained by solving for the (s,z) which satisfy the equilibrium condition. The betting pool and bookie profits for the spread-odds (s,z) game with (s,z) in V then follows directly.

Proposition 1: For arbitrary (not necessarily equilibrium) (s,z),

$$\Lambda^R = n\int_{w>T+A} \lambda(w - T - A)dM(w),$$

$$\Lambda^B = n\int_{w<T-A} \lambda(-w + T - A)dM(w),$$

where $T = s - \mu_M - [Q(z + c/2) + Q(z - c/2)]/2$ and $A = [Q(z + c/2) - Q(z - c/2)]/2$.

Proof: This follows from the assumed betting criterion, the representation of market beliefs, and the direct substitution $w = \mu - \mu_M$.

Proposition 2: The equilibrium set is $V = \{(s,z) | s(z) = \mu_M + [Q(z + c/2) + Q(z - c/2)]/2\}$.

Proof: By the symmetry of M we have

$$\Lambda^R - \Lambda^B = n\int_{w>A+T} \lambda(w - T - A)dM - n\int_{w>A-T} \lambda(w + T - A)dM.$$

This expression is zero at $T = 0$ and, using the assumption that M is increasing, it may be verified that $\Lambda^R - \Lambda^B < 0$ for $T > 0$ and $\Lambda^R - \Lambda^B$

> 0 for $t < 0$. Hence $\Lambda^R - \Lambda^B = 0$ if and only if $T = 0$, and this is the proposition.

Proposition 3: Bookie profits for $[s(z), z]$ in V are

$$c(\Lambda^R + \Lambda^B) = cn\int_{w>A} \lambda(w - A)dM(w).$$

An optimal game is therefore an $[s^*(z^*), z^*]$ such that $A(z^*:c) = \min_z A(z:c)$.

Proof: The first part of the proposition follows on substituting $s(z)$ for s in proposition 1 and then using the symmetry of M. The second part follows on noting that with M increasing the expression for profits increases as A decreases.

Substituting $s(z)$ for s in the previous expression for nonbettors, we find that as (s,z) varies over V, the nonbettors are those individuals whose μ_i fall in the interval $[\mu_M - A(z:c), \mu_M + A(z:c)]$. Since the μ_i are symmetric about μ_M the proportion of potential bettors who purchase no bets is $1 - 2M[A(z:c)]$. Since M is increasing this proportion is minimized at $\min_z A(z:c)$, which is the same as the maximum-profit condition. The objectives of maximizing profits and maximizing the participation of bettors in the market are therefore equivalent, given our assumptions on beliefs and betting behavior.

When $A(z:c)$ is a constant function of z, all betting games will yield equal profits and an equal number of bettors. This occurs when Q is linear (for $0 < p < 1$), and this in turn implies that G is uniform, the example of the previous section. Conversely, when G is not uniform A is not constant, and some betting games will not be produced because they generate lower profits. The bookie fee in effect drives a wedge, consisting of the nonbettors, between Rams and Bears bettors, and the size of this wedge depends on which spread-odds (s,z) game is produced.

Finally, consider the point-spread game with the equilibrium spread μ_M. If Rams bettors compute an odds game at which their behavior is identical with the point-spread game, they will find $z_R = G\{Q[(1 - c)/2] - \mu_M\} + c/2$. The Bears bettors will compute $z_B = G\{Q[(1 + c)/2] - \mu_M\} - c/2$. If G is not uniform the equilibrium odds are not equal to z_R or z_B so that the odds game with $z = z_R$ or z_B is not a feasible game for the bookie. The existence of equalizing odds for any given bettor does not therefore imply that all betting games are equivalent. The point-spreads versus odds question depends on betting behavior not at equalizing odds but at the equilibrium odds, and unless G is uniform these two odds will not be equal.

Given the assumed beliefs and betting criterion, proposition 3 gives the optimality condition for spread-odds (s,z); namely, the bookie

produces an $[s(z), z]$ game such that $A(z:c)$ is a minimum. The next section translates this condition into simple conditions on G.

Remark

Let the (median) consensus probability of the event $X < s$ be denoted by $\pi_M(s)$ where half of the individuals in N believe $P(X < s) < \pi_M(s)$ and half believe $P(X < s) > \pi_M(s)$. Suppose further that this consensus is correct: If $\pi_M(s) = z$ for a large number of football games, then z is the frequency of $X < s$. It may be verified that for the present model $\pi_M(s) = G(s - \mu_M)$. Furthermore, if $c = 0$ everyone in N purchases a wager (the set of nonbettors has probability zero), and π_M is revealed by the equilibrium set since $V = [(s,z) \mid \pi_M(s) = z]$. However, from proposition 2 we see that with $c > 0$ there will exist discrepancies between $\pi(s) = z$, (s,z) in V and the true measure of the market's opinion $\pi_M(s)$. This discrepancy arises because with $c > 0$ some potential bettors do not wager, and their opinions are not fully represented by the equilibrium set. For example, if $c = .2$, G is a mean zero, variance 49 Gaussian distribution and $\mu_M = 7$, then one finds the equilibrium odds is $\bar{z} = .179$, and this is a biased estimate of the potential market's belief regarding $X < 0$ since $\pi_M(0) = .159$. Hence $\bar{z}$ will not reflect the frequency of $X < 0$ even if the consensus beliefs of potential bettors are correct. Such discrepancies might be a reason for the observed difference between the probabilities implied by the odds in horse race betting and the frequency with which horses win (see Ali [1977, p. 810] for a similar result; see also Seligman [1975]).

VI

Suppose in addition to previous assumptions that G is strictly unimodal; that is, $g(0) > g(x)$ for all x and $g(x)$ increasing (decreasing) as x is negative (positive). With this assumption μ_i is the expectation, median, and mode or most likely outcome of X. In such a world we have:

Theorem 1: If G is strictly unimodal, then point spreads is the unique optimal betting game.

Proof: The first-order condition for the minimum of A is $q(z + \frac{1}{2}c) - q(z - c/2) = 0$, or $g[Q(z + c/2)]^{-1} - g[Q(z - c/2)]^{-1} = 0$. With G symmetric and strictly unimodal this holds if and only if $Q(z + c/2) = -Q(z - c/2) = Q(1 - z + c/2)$, or $z = \frac{1}{2}$. One may now verify that A is decreasing for $z < \frac{1}{2}$ and increasing for $z > \frac{1}{2}$ so that at $z = \frac{1}{2}$, A is at its unique minimum value. By proposition 3, $z = \frac{1}{2}$, or the point-spread game is the unique optimal betting game.

To obtain simple necessary and sufficient conditions which characterize the optimal game, suppose c is small and use a linear

approximation for $Q(z + c/2)$. Expanding around $c = 0$ and suppressing quadratic and higher-order terms gives $Q(z + c/2) = Q(z) + q(z)c/2$. Up to this order of approximation $s(z) = \mu_M + Q(z)$, $\bar{z} = G(-\mu_M) = 1 - G(\mu_M)$, and $A(z:c) = q(z) = g[Q(z)]^{-1}$.

Theorem 2: For c small, spread-odds $[s(z^*),z^*]$ is an optimal game if and only if $Q(z^*)$ is a global mode of $G(x)$.[7]

Proof: Minimizing $A(z:c)$ with c small is approximately the same as minimizing $q(z) = g[Q(z)]^{-1}$, and by definition this minimum is achieved at a mode of $G(x)$. (Notice that if $Q[z^*]$ is a mode, then by symmetry so is $-Q[z^*] = Q[1 - z^*]$, so that $[s(1 - z^*), 1 - z^*]$ is also an optimal game.)

Theorem 3: For c small, point spreads beat odds if and only if $g(0) > g(\mu_M)$.

Proof: With c small, $A(\bar{z} : c) = g(\mu_M)^{-1}$ (since $g[\mu_M] = g[-\mu_M]$) and $A(\frac{1}{2}:c) = g(0)^{-1}$ so that the theorem follows from proposition 3.

It should be emphasized that the simple optimality conditions in this section are a consequence of the simple specification of beliefs and betting behavior. Under less restrictive assumptions there will undoubtedly exist different configurations of beliefs, betting behavior, and preferences which are revealed by the bookie's production of point spreads.

The conditions illustrate how market beliefs and betting behavior can be mapped via the bookie into a particular betting game even when betting games are considered equally thrilling. For example, under the assumptions of theorem 2, the existence of point-spread betting does not reveal bettor preferences but instead reveals an aspect of a bettor's beliefs, namely, that $0 = Q(\frac{1}{2})$ is the mode of G. There is not a one-to-one correspondence between point spreads and the previously described bettor-preferences explanation, and such an explanation therefore is not logically necessary in order to understand the existence of point-spread betting.

VII

In this section the location-family-of-market-beliefs assumption is dropped, individual betting is supposed consistent with the positive *EMV* requirement but is otherwise arbitrary, and a very simple example is presented in which point spreads versus odds is no contest. Only the point-spread and odds games are compared, and there is no attempt to obtain simple or general optimality conditions. The example is presented to illustrate how point spreads can be favored over odds when bettors use an arbitrary *EMV*-based decision criterion. The

[7] x^* is a global mode of G if $g(x^*) \geq g(x)$ for all x.

more specific criterion used in the previous sections, while influencing the particular optimality conditions, is not therefore crucial for the bookie to favor one betting game over another.

Suppose N consists of just two potential bettors. Let $c = 1/9$ and suppose beliefs are given by the uniform distributions,

$$F_1(x) = \frac{5}{18}, \qquad \frac{5}{3} \leq x \leq \frac{13}{3},$$

$$F_2(x) = \frac{x}{12} + \frac{2}{9}, \qquad -\frac{8}{3} \leq x \leq \frac{28}{3}.$$

These distributions are depicted in figure 1. Observe that the distributions do not differ by just location and are not therefore covered by the previous sections. Also note that there are disparate beliefs both in the sense that $F_1 \neq F_2$ and also in the sense that beliefs regarding the event $X < 0$ are much less dispersed than beliefs regarding, say, $X < 2$.

The *EMV*s of the Rams and Bears bets for both individuals given the point-spread game are graphed in figure 2 as a function of s. From the figure one sees that when $4/3 \leq s \leq 24/9$, individual 1 is a Bears bettor and individual 2 is a Rams bettor. The equilibrium point spread will therefore be somewhere in this interval, with the particular spread being dependent on each individual's betting criterion. Under the criterion of Section III the equilibrium spread is 2; if individuals purchase a quantity of wagers which is proportional to the *EMV* of the wager, then $\bar{s} = 16/9$. But in any event both potential

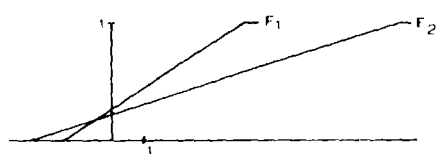


FIG. 1.—Beliefs of potential bettors 1 and 2

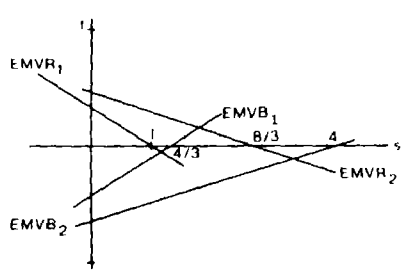


FIG. 2.—*EMV*s of Bears and Rams bets for potential bettors given point-spread game

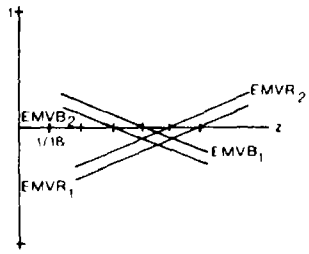


FIG. 3.—*EMV*s of Bears and Rams bets for potential bettors given odds game

bettors will participate in the betting market by purchasing some point-spread bets.

The *EMV*s of the Rams and Bears bets given the odds game are graphed in figure 3 as a function of z. For $z < 2/9$ the *EMV* of a Bears bet is positive for both individuals and for $z > 5/18$ the *EMV* of a Rams bet is positive for both individuals; hence, the equilibrium odds cannot be in either of these intervals. The only way to satisfy the equilibrium condition is to let $2/9 \leq z \leq 5/18$, and with any such odds both individuals purchase no bets. Therefore point spreads versus odds is no contest. If the point-spread device had not been invented and wagers could only be made on the winner of the game, this football betting market would not exist.

VIII

So what is the answer to the point-spreads versus odds question? Perhaps the answer is that point spreads are more fun than odds. Since this, however, is not the only logical possibility, the answer ultimately must be based on empirical evidence. In a world where point spreads are not considered more fun than odds, point spreads can exist as the game produced by a profit-maximizing bookie. Difference of opinion, besides leading to the formation of a betting market, can also have implications for the particular betting game produced by the bookie. Finally, to profit from these findings you should look for an informal setting in a location-family-of-beliefs world where bettors use a criterion as described in Section III. If a profit-maximizing bookie is producing the point-spread game, you might then wager that your fellow bettors have expectations equal to their modes.

References

Ali, Mukhtar M. "Probability and Utility Estimates for Racetrack Bettors." *J.P.E.* 85, no. 4 (August 1977): 803–16.

Gruen, Arthur. "An Inquiry into the Economics of Race-Track Gambling." *J.P.E.* 84, no. 1 (February 1976): 169–77.

Pankoff, Lyn D. "Market Efficiency and Football Betting." *J. Bus.* 41 (April 1968): 203–14.

Seligman, Daniel. "A Thinking Man's Guide to Losing at the Track." *Fortune* 92 (September 1975): 81–87.

Smith, Vernon L. "Economic Theory of Wager Markets." *Western Econ. J.* 9 (September 1971): 242–55.

Winkler, Robert L. "Probabilistic Prediction: Some Experimental Results." *J. American Statis. Assoc.* 66 (December 1971): 675–85.

Time Preference and International Lending and Borrowing in an Overlapping-Generations Model

Willem H. Buiter

University of Bristol and NBER

Two economies, represented by Diamond-type overlapping-generations models and differing only in their pure rates of time preference, are joined together. Capital formation, balance-of-payments behavior, and welfare are compared under autarky and openness. With a positive natural rate of growth, the low-time-preference country runs a current account surplus in the steady state but not necessarily outside it. If preexisting capital is not shiftable between countries, integration in the world economy makes the high-time-preference country worse off in the short run. The ranking of stationary utility levels under autarky and openness is ambiguous.

I. Introduction

The purpose of this paper is to explain international capital movements from differences in time preference and to evaluate the short-run and long-run welfare implications of a change from a situation of trade and financial autarky to one of openness in trade and finance. The analysis is conducted in terms of a deterministic model of two countries, each one of which consists of a Samuelson-Diamond type of overlapping-generations model and produces one identical good. All

I would like to thank Jeff Carmichael, Jon Eaton, Smith Freeman, and Paul Krugman for useful comments. Two anonymous referees suggested significant alterations in an earlier version of the paper that were very helpful. The research reported here is part of the National Bureau of Economic Research's Program in International Studies. Any opinions expressed are those of the author and not those of the NBER.

Journal of Political Economy, 1981, vol. 89, no. 4)
© 1981 by The University of Chicago. 0022-3808/81/8904-0006$01.50

determinants of international trade and international lending and borrowing other than a single taste parameter, the pure rate of time preference, are ignored. There are no differences in technology, in exogenous factor endowments, in scale, and in market structure; nor are there taxes, subsidies, and other man-made wedges between buyers' and sellers' prices or between domestic and foreign prices.

There is an extensive literature on the determinants of international capital movements. Starting from the accounting identity that a country's net international capital outflow equals the difference between national saving and domestic capital formation, a number of different approaches have been taken to the determinants of saving and investment. Ad hoc saving functions, mainly specifying saving as a constant fraction of disposable income, are the most common (Borts 1964; Neher 1970; Fischer and Frenkel 1972; Onitsuka 1974; Hori and Stein 1977; Ruffin 1979). Stiglitz (1970) derived saving behavior from the utility-maximizing choices of infinite-lived individuals. The approach adopted here and in Fried (1980) has optimizing, overlapping generations with finite lives. This is especially useful for the short-run and long-run welfare analyses of a move from autarky to openness. Capital formation is carried to the point where the marginal product equals the interest rate. This is different from the exogenous investment of Neher (1970) and Onitsuka (1974), the ad hoc (Keynesian) investment function of Borts (1964), and the cost-of-adjustment investment function of Fischer and Frenkel (1972). Only a single good, which can be consumed or used as a capital good, is produced in both countries. This is the same as in Neher (1970), Onitsuka (1974), and Ruffin (1979). Stiglitz (1970) models production of and trade in distinct consumer goods and capital goods, as do Fischer and Frenkel (1972). Hori and Stein (1977) have distinct consumer goods and capital goods, but each of the two countries in their model is assumed to specialize completely in production of one of these goods. Borts (1964) distinguishes between traded and nontraded (labor-intensive) goods; traded goods are imports and exports; the traded goods can be either consumer goods or investment goods. The one-commodity model cannot be used to analyze the effect of capital mobility on the terms of trade, a concern of Borts (1964), Stiglitz (1970), and Hori and Stein (1977). The one-commodity structure also means that trade in goods must be matched by international financial mobility. Stiglitz (1970) has trade in consumer goods and capital goods but no financial capital mobility. Fischer and Frenkel (1972) have international trade in capital goods, in consumer goods, and in financial claims—as does Borts (1964), although there all capital goods are imported. Finally, this paper considers a two-country model, as do Stiglitz (1970), Hori and Stein (1977), and Ruffin

(1979). The small-country case was analyzed by Borts (1964), Neher (1970), Fischer and Frenkel (1972), and Onitsuka (1974). In a recent paper, Findlay (1978) has approached the problem of international trade, time preference, and interest equalization from an "Austrian" viewpoint. This paper can be viewed in part as a neoclassical complement to Findlay's analysis. Section II describes the autarky equilibrium of the two countries. Section III analyzes the open, integrated world economy.

One of the conclusions of the positive analysis, that in the steady state (but not necessarily outside it) the high-time-preference country runs a current account deficit, is hardly surprising and deserves mentioning mainly because of its rigorous foundation in optimizing behavior. The welfare analysis of autarky versus openness in Section IV presents some rather less conventional results. If preexisting capital is nonshiftable between countries, the high-time-preference country is worse off under openness in the short run. The long-run welfare analysis is more complex but also fails to generate unambiguous welfare improvements for both countries under openness.

II. The Autarky Equilibrium

The world economy consists of two countries, the "home" country and the "foreign" country, identical in every respect except in taste. Each country is represented by competitive output and factor markets, two overlapping generations, and an identical, well-behaved constant returns to scale production function f, a model due to Diamond (1965).[1]

Notation

c_t^1 = consumption while young by a member of generation t;
c_t^2 = consumption while old by a member of generation t;
L_t = size of generation t;
K_t = capital stock in period t;
k_t = K_t/L_t;
w_t = real wage in period t;
r_t = interest rate on savings carried from period $t - 1$ into period t;
n = one-period proportional rate of growth of population $L_{t+1} = (1 + n)L_t$; $n \geq 0$;
ρ = pure rate of time preference; $\rho \geq 0$;
A_t = wealth in period t;

[1] The original development of the overlapping-generations model is by Samuelson (1958). See also Cass and Yaari 1966.

$a_t = A_t/L_t$;
B_t = trade balance surplus of the home country in period t;
$b_t = B_t/L_t$;
G_t = current account surplus of the home country in period t;
$g_t = G_t/L_t$.

Momentary Equilibrium

The momentary or single-period equilibrium of the closed economy is summarized in equations (1)–(7):

$$\max_{c_t^1, c_t^2} u(c_t^1, c_t^2), \tag{1}$$

$$w_t - c_t^1 \geqslant c_t^2(1 + r_{t+1})^{-1}, \tag{2}$$

$$c_t^1, c_t^2 \geqslant 0, \tag{3}$$

$$w_t = f(k_t) - k_t f'(k_t), \tag{4}$$

$$r_t = f'(k_t), \tag{5}$$

$$w_t - c_t^1 = k_{t+1}(1 + n), \tag{6}$$

$$k_t \geqslant 0. \tag{7}$$

The utility function, u, is twice differentiable, strictly quasi-concave, and increasing in c^1 and c^2. The quantity $u_1(0, c^2) = u_2(c^1, 0) = +\infty$; $u_1(\infty, c^2) = u_2(c^1, \infty) = 0$. The function f is twice differentiable; $f(0) = 0$; $f' > 0$; $f'' < 0$; $f'(0) = +\infty$ and $f'(\infty) = 0$.

Individuals within a given country are identical within and across generations. People live for two periods, work in the first period of their lives, and retire in the second. Labor supply is inelastic. The individual's optimization problem is given in equations (1), (2), and (3). The utility of lifetime consumption is maximized subject to the lifetime budget constraint (2). The conditions on u ensure that the budget constraint will hold with equality and that an interior solution will be obtained for c_t^1 and c_t^2. Equations (4) and (5) state that the labor market and capital rental market are competitive. The capital stock at the beginning of period $t + 1$ equals the value of the saving in period t. This economy-wide capital market equilibrium condition is given in equation (6). Real capital is the only store of value. All saving in period t is performed by the young members of generation t. The old dissave previously accumulated wealth. The conditions on f ensure that k_t will be positive.

When we consider the consequences of different pure rates of time preference, the utility function $u(c_t^1, c_t^2)$ will be specialized to the

following additively separable form:

$$u(c_t^1, c_t^2) = v(c_t^1) + (1 + \rho)^{-1} v(c_t^2). \tag{8}$$

Here ρ is the (constant) pure rate of time preference. More generally, the pure rate of time preference is the marginal rate of substitution between consumption when young and consumption when old, when equal amounts are consumed in both periods, minus one. It can be a function of the level of consumption: $\rho(c) = [u_1(c, c)/u_2(c, c)] - 1$.

The general solution of (1)–(7) is straightforward and can be found, for example, in Diamond (1965). The interior first-order conditions for the individual are:

$$\frac{u_1(c_t^1, c_t^2)}{u_2(c_t^1, c_t^2)} = 1 + r_{t+1}. \tag{9}$$

For the utility function of equation (8), this becomes:

$$\frac{v'(c_t^1)(1 + \rho)}{v'(c_t^2)} = 1 + r_{t+1}. \tag{9'}$$

The consumer's intertemporal equilibrium is graphed in figure 1. Equations (9) and (2) (the latter holding with equality) can be solved for c_t^1 and c_t^2 as functions of w_t and r_{t+1}:

$$c_t^1 \equiv c^1(w_t, r_{t+1}), \tag{10a}$$

$$c_t^2 = c^2(w_t, r_{t+1}).^2 \tag{10b}$$

Similarly, equations (9′) and (2) can be solved to yield:

$$c_t^1 = c^1(w_t, r_{t+1}, \rho), \tag{10a'}$$

$$c_t^2 = c^2(w_t, r_{t+1}, \rho). \tag{10b'}$$

If consumption in both periods is a normal good, $1 > c_{w}^1, c_{w}^2 > 0$. We shall assume this to be the case.

$$^2\frac{\partial c_t^1}{\partial w_t} = \frac{-[u_{12} - (1 + r_{t+1})u_{22}]}{[u_{11} - (1 + r_{t+1})u_{21}](1 + r_{t+1})^{-1} - u_{21} + (1 + r_{t+1})u_{22}},$$

$$\frac{\partial c_t^1}{\partial r_{t+1}} = \frac{u_2 - c_t^2(1 + r_{t+1})^{-1}[u_{21} - (1 + r_{t+1})u_{22}]}{u_{11} - (1 + r_{t+1})[2u_{21} - (1 + r_{t+1})u_{22}]}$$

$$\frac{\partial c_t^2}{\partial w_t} = (1 + r_{t+1})\left(1 - \frac{\partial c_t^1}{\partial w_t}\right),$$

$$\frac{\partial c_t^2}{\partial r_{t+1}} = -(1 + r_{t+1})\frac{\partial c_t^1}{\partial r_{t+1}} + \frac{c_t^2}{1 + r_{t+1}}.$$

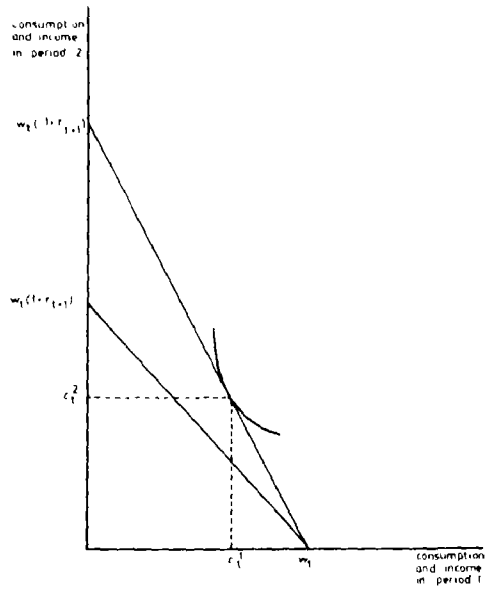


FIG. 1.—The consumer's intertemporal equilibrium

Steady-State Equilibrium

The steady-state or long-run equilibrium is that sequence of momentary equilibria in which the lifetime per capita consumption profile remains constant from generation to generation. It can be characterized by the following set of equations:

$$\frac{u_1(c^1, c^2)}{u_2(c^1, c^2)} = 1 + r, \tag{11}$$

$$\frac{v'(c^1)(1 + \rho)}{v'(c^2)} = 1 + r, \tag{11'}$$

$$w - c^1 = \frac{c^2}{1 + r}, \tag{12}$$

$$w = f(k) - kf'(k), \tag{13}$$

$$r = f'(k), \tag{14}$$

$$w - c^1 = k(1 + n). \tag{15}$$

By substituting equations (13) and (14) into (12) and (15), we obtain two equations that represent the stationary decentralized (or competitive) consumption possibility locus:

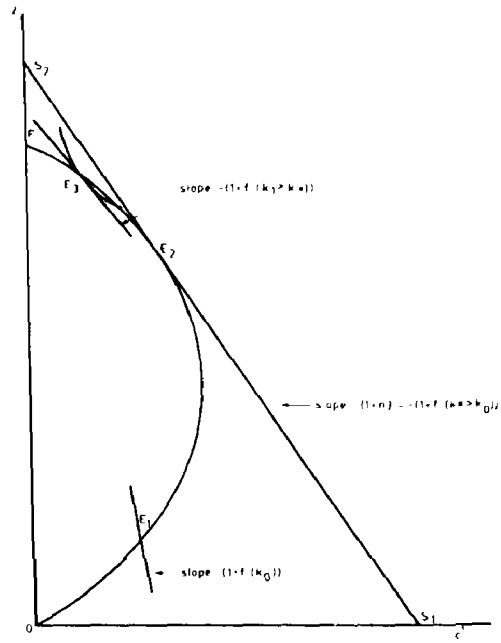


FIG. 2.—The stationary competitive possibility locus

$$f(k) - kf'(k) = c^1 + \frac{c^2}{1 + f'(k)}, \tag{16a}$$

$$f(k) - kf'(k) = c^1 + k(1 + n). \tag{16b}$$

These two equations can be solved for c^1 as a function of c^2:

$$c^1 = \psi(c^2);\ \psi' = -\frac{(1 + n + kf'')}{(1 + n)(1 + f' + kf'')}. \tag{17}$$

For the Cobb-Douglas production function $f(k) = k^\alpha$, $0 < \alpha < 1$, the stationary decentralized consumption possibility locus has the shape of OF in figure 2.[3] Along this locus: $(dc^1/dc^2) = -\{[1 + n + (\alpha - 1)\alpha k^{\alpha-1}]/(1 + n)(1 + \alpha^2 k^{\alpha-1})\}$. At the origin (with $k = 0$), $(dc^2/dc^1) = [\alpha/(1 - \alpha)](1 + n) > 0$. The capital-labor ratio, k, increases monotonically as one moves from O to F. As k approaches infinity (which would

[3] The historical origins of fig. 2 are not completely clear. It was a well-established part of the "Yale blackboard tradition" when I came to Yale as a graduate student in 1971. James Tobin suggests that David Cass may have been the first one to utilize it. It has since been used as an expository device by Ihori (1978), Buiter (1979), and Carmichael (1979).

be for $c^1 < 0$), (dc^2/dc^1) becomes -1. When $c^1 = 0$ at F, $(dc^2/dc^1) = -[1 + \alpha^2(1 + n)(1 - \alpha)^{-1}](1 - \alpha)^{-1} < 0$. At the golden-rule capital-labor ratio, defined by $f'(k) = n$, (dc^2/dc^1) is $-[1 + f'(k)] = -(1 + n)$.[4] I shall assume in what follows that OF is as in figure 2: strictly concave toward the c^2 axis for nonnegative values of c^1 and c^2. This is certainly the case for the Cobb-Douglas production function but need not be true for alternative, otherwise well-behaved neoclassical production functions. It is, for example, quite possible for OF, while remaining strictly concave toward the origin, to be upward sloping at F. This would reflect extreme overaccumulation of capital: An increase in k at high values of k depresses the marginal product of capital to such an extent that reductions in both c^1 and c^2 are required to generate the additional saving required to maintain the higher value of k.

A stationary competitive equilibrium is described by equations (16a) and (16b) and a tangency of a private budget constraint with slope $1 + f'(k) = 1 + r$ to an indifference curve, as given in (11) or (11'). At capital-labor ratios above the golden-rule level, the budget constraint cuts the OF curve from above if the production function is Cobb-Douglas.[5] In what follows I shall assume that this condition holds.

In figure 2 I have drawn such an equilibrium at E_3, corresponding to a capital-labor ratio in excess of the golden-rule capital-labor ratio. This stationary competitive equilibrium is inefficient in the sense that any lower capital-labor ratio that does not lie below the golden-rule capital-labor ratio can sustain higher stationary paths of per capita consumption and does not require any sacrifice in consumption to move to these higher stationary paths if only physical resource constraints are taken as binding. The physical resource constraint faced by an omniscient and omnipotent social planner is $c^2_{t-1}L_{t-1} + c^1_t L_t \leq f(k_t)L_t + K_t - K_{t+1}$, or

$$c^2_{t-1}(1 + n)^{-1} + c^1_t \leq f(k_t) + k_t - k_{t+1}(1 + n). \tag{18}$$

The stationary resource constraint is:

[4] In what follows I shall assume that the golden-rule capital-labor ratio is less than the capital-labor ratio at F. This need not be the case. From (16b) it is clear that with a Cobb-Douglas production function, $c^1 = 0$ implies either $k = 0$ (at the origin) or $(1 - \alpha)k^{\alpha-1} = (1 + n)$. (This requires that k be less than unity at F.) Thus, at F, $k^{\alpha-1} = (1 + n)/(1 - \alpha)$. At the golden rule, $k^{\alpha-1} = n/\alpha$. It is likely, but not inevitable, that k at F exceeds the golden-rule capital-labor ratio.

[5] The slope of the consumption possibility frontier, S_1, is given by: $S_1 = [(1 + n)(1 + \alpha^2 k^{\alpha-1})]/[(1 - \alpha)\alpha k^{\alpha-1} - (1 + n)]$. The slope of the budget constraint, S_2, is given by $S_2 = -(1 + \alpha k^{\alpha-1})$. When $f'(k) = \alpha k^{\alpha-1} < n$, both are negative.

$$S_1/S_2 = \frac{1 + n + (1 + n)\alpha^2 k^{\alpha-1}}{1 + n + (1 + n)\alpha^2 k^{\alpha-1} + (1 - \alpha)\alpha k^{\alpha-1}(n - \alpha k^{\alpha-1})}.$$

This is < 1 if $n > \alpha k^{\alpha-1}$. I am obliged to an anonymous referee for saving me from error here.

$$\frac{c^2}{1+n} + c^1 \leq f(k) - nk. \tag{19}$$

A social planner aiming to maximize the stationary value of per capita resources available for consumption by the young and the old chooses k such that $f'(k) = n$. The social planner's stationary consumption possibility locus is $S_1 - S_2$ in figure 2. The decentralized competitive economy will settle somewhere on OF if the economy has a stable equilibrium.[6] An equilibrium such as E_1 in figure 2 represents a steady-state capital-labor ratio below the golden rule. This cannot be characterized as inefficient because of (temporary) sacrifices in consumption that would be required to reach a higher capital-labor ratio.

Stability

Substituting (10a) into (6) and using (4) and (5), we obtain a first-order difference equation in k_t that describes the evolution of the model from arbitrary initial conditions:

$$f(k_t) - k_t f'(k_t) - c^1[f(k_t) - k_t f'(k_t), f'(k_{t+1})] = k_{t+1}(1+n). \tag{20}$$

The model will be locally stable when $|(dk_{t+1}/dk_t)| < 1$, that is, when

$$|(c^1_w - 1)kf''(1 + n + c_r f'')^{-1}| < 1. \tag{21}$$

This says that a higher value of the capital stock in period t, $k_t L_t$, should not boost saving by young members of generation t by more than the increase in $k_t L_t$. Note that the assumption that goods are normal ($0 < c^1_w < 1$) and the assumption that an increase in the interest rate reduces current consumption ($c^1_r < 0$) are insufficient to ensure stability.

Properties of the Autarky Equilibrium

We can now establish five propositions about the autarky equilibria of two economies that are identical in all respects except for their pure rates of time preference. The utility function is given in equation (8).

Proposition 1.—Given w_t and r_{t+1}, a consumer with a higher pure rate of time preference will choose a higher value of consumption when young, c^1_t.

Proof.—We want to show that the consumption function $c^1_t = c^1(w_t, r_{t+1}, \rho)$ has the property $c^1_\rho = (\partial c^1/\partial \rho) > 0$. Differentiating $\{(1 + \rho)v'(c^1)/v'[(w - c^1)(1 + r)]\} = 1 + r$, we obtain

[6] The assumptions we have made so far suffice to ensure the existence of a long-run equilibrium (Carmichael 1979). A (probably overly strong) sufficient condition for the uniqueness of the steady-state equilibrium is that preferences be homothetic (Carmichael 1979).

$$\frac{\partial c^1}{\partial \rho} = \frac{v'(c^1)}{1+\rho}\left\{v''(c^1) + \left[(1+\rho)\frac{v'(c^1)}{v'(c^2)}\right]^2 \frac{v''(c^2)}{1+\rho}\right\}^{-1}.$$

The utility function is increasing in c^1 and c^2, so $v' > 0$. Strict convexity of the indifference curve requires $v'' < 0$. Therefore $(\partial c^1/\partial \rho) > 0$.

Proposition 2.—When one starts from a given k_t (and therefore a given w_t), a higher pure rate of time preference is associated with a lower value of k_{t+1} (and a higher value of c_t^1) if a higher interest rate does not reduce saving.

Proof.—For the case of the additively separable utility function, equation (20) can be rewritten as: $f(k_t) - k_t f'(k_t) - c^1[f(k_t) - k_t f'(k_t), f'(k_{t+1}), \rho] = k_{t+1}(1+n)$. Thus $(\partial k_{t+1}/\partial \rho) = \{-c_\rho^1/[1 + n + c_r^1 f''(k_{t+1})]\}$. A sufficient, but not necessary, condition for $(\partial k_{t+1}/\partial \rho)$ to be negative is $c_r^1 < 0$.

Proposition 3.—Given ρ, a lower value of k_t will be associated with a lower value of k_{t+1} if consumption in period 2 is a normal good and a higher interest rate does not reduce saving.

Proof.—From (20) we obtain $(\partial k_{t+1}/\partial k_t) = \{-[(1 - c_w^1)k_t f''(k_t)]/[1 + n + c_r^1 f''(k_{t+1})]\}$. This will be positive if $c_w^1 < 1$ and $c_r^1 \leq 0$. (Note that $c_w^1 < 1 \Leftrightarrow c_w^2 > 0$.)

Propositions 2 and 3 immediately imply proposition 4.

Proposition 4.—Consider two economies, identical in all respects except the pure rate of time preference, which is higher for the home country than for the foreign country. When one begins from any common initial capital-labor ratio at $t = t_0$, the capital-labor ratio of the high-time-preference country will be below that of the low-time-preference country for all $t > t_0$ if consumption in period 2 is a normal good and a higher interest rate does not reduce saving.

Proposition 4 tells us something about the behavior of the two economies were they to start from the same initial condition. It might still seem possible that if the high-pure-rate-of-time-preference country were to start off with an initial capital-labor ratio, k_{t_0}, sufficiently above the initial capital-labor ratio of the low-pure-rate-of-time-preference country, $\bar{k}_{t_0}$, k_t could remain above $\bar{k}_t$ indefinitely. To show that if $k_{t_0} > \bar{k}_{t_0}$ the two paths of the capital-labor ratio $\{k_t\}$ and $\{\bar{k}_t\}$ will cross, it is necessary to show that the steady-state capital-labor ratio of the high-time-preference country, k, is less than the steady-state capital-labor ratio of the low-time-preference country, $\bar{k}$, and that the model is globally stable. Proposition 5 states the conditions under which the steady-state capital-labor ratio is a decreasing function of the pure rate of time preference. Sufficient conditions for global stability are given in Carmichael (1979). After

the two paths cross for the first time, they never cross again because of propositions 2 and 3.

Proposition 5.—Under autarky, the country with the higher pure rate of time preference will have the lower steady-state capital-labor ratio if the model is locally stable, consumption in both periods is a normal good, and an increase in the interest rate lowers consumption in period 1.[7]

Proof.—Consider again the consumption function $c_t^1 = c^1(w_t, r_{t+1}, \rho)$. From proposition 1 we know that $c_\rho^1 > 0$. We find the effect of a higher value of ρ on steady-state k from $f(k) - kf'(k) - c^1 = k(1 + n)$. Differentiating this yields: $(\partial k/\partial \rho) = c_\rho^1[(c_w^1 - 1)kf'' - (c_r^1 f'' + 1 + n)]^{-1}$. With $c_\rho^1 > 0$, $(\partial k/\partial \rho) < 0$ requires $(c_w^1 - 1)kf'' - (c_r^1 f'' + 1 + n) < 0$. This is ensured by stability (eq. [21]) if $0 < c_w - 1$ and $c_r^1 \leq 0$.

III. The Two-Country Equilibrium

The two countries that will be linked together in an international commodity market and an international capital market are identical in all respects, except for the pure rate of time preference. The home country is assumed to have a higher rate of pure time preference than the foreign country. For the home country, we adopt the notation developed for the autarkic economy. All variables associated with the foreign country are distinguished by a tilde above the relevant variable. Comparing the two countries, we have:

$$v = \tilde{v}, \tag{22a}$$

$$f = \tilde{f}, \tag{22b}$$

$$n = \tilde{n}, \tag{22c}$$

$$L_t = \tilde{L}_t,^{8} \tag{22d}$$

$$\rho > \tilde{\rho}. \tag{22e}$$

Output can be moved costlessly between the two countries.

In this "pure-absorption" model,[9] international trade and international lending and borrowing (international capital mobility) are part and parcel of the same transaction. In a one-commodity model, the only way to pay for an extra unit of output today is with a promise of future output. Each trade balance transaction has to involve credit.

[7] Since only local stability is assumed, proposition 5 is valid only for small differences between the two countries' pure rates of time preference.

[8] The model is easily generalized to include scale differences, as long as the natural rate of growth in both countries is the same.

[9] I owe this description to Bill Branson.

There can be no current quid pro quo. We make the convenient assumption of perfect international mobility of financial capital: Ownership claims to domestic and foreign real capital are perfect substitutes in private portfolios in both countries. This means that interest rates will be equalized in the world economy:

$$r_t = \tilde{r}_t \text{ for all } t. \tag{23}$$

There is a single perfect world financial capital market and a single perfect world market for the current production of the homogeneous commodity that can be used as a consumption good or a capital good.[10] With international capital mobility, it is essential to distinguish between the capital stock used in production in the home country, K_t, and the value of claims on real capital, domestic or foreign, owned by domestic residents, A_t. In the national income accounts the corresponding distinction is between domestic income or product and national income or product. Location and ownership of physical capital no longer coincide. From interest equalization, free trade in capital goods, and identical, linear homogeneous production functions, it follows that capital-labor ratios and wage rates are equal in the two countries:

$$k_t = \tilde{k}_t, \tag{24a}$$

$$w_t = \tilde{w}_t. \tag{24b}$$

The behavior of the integrated world economy is given by the optimizing behavior of domestic and foreign households, factor market equilibrium conditions, and the world capital market equilibrium condition. Domestic household behavior is given by equations (1) (or its additively separable version [8]), (2), and (3). Foreign household behavior is represented by the same functions, with $\tilde{c}_t^1$, $\tilde{c}_t^2$, and $\tilde{\rho}$ substituted for c_t^1, c_t^2, and ρ, respectively. Factor market equilibrium in both countries is given by (4) and (5), the world capital market equilibrium condition by

$$w_t - \frac{1}{2}(c_t^1 + \tilde{c}_t^1) = k_{t+1}(1 + n). \tag{25}$$

Openness is reflected in the equalization of interest rates and of marginal products of capital. Equation (25) states that world saving in period t, $(w_t - c_t^1)L_t + (\tilde{w}_t - \tilde{c}_t)\tilde{L}_t$, equals the world capital stock in period $t + 1$, $K_{t+1} + \tilde{K}_{t+1} = A_{t+1} + \tilde{A}_{t+1}$.

The individual optimization programs yield, in the additively

[10] When the short-run gains from international economic integration are considered, it is assumed that the preexisting national capital stocks cannot be relocated. Newly produced capital goods can be located costlessly in either country.

separable case, equation (9′) as the first-order condition for the home country. Replacing c_t^1, c_t^2, and ρ by $\bar{c}_t^1$, $\bar{c}_t^2$, and $\bar{\rho}$, (9′) also represents the first-order condition for the foreign country. Given the predetermined value of k_t ($= \bar{k}_t$), the model determines the momentary equilibrium values of c_t^1, c_t^2, $\bar{c}_t^1$, $\bar{c}_t^2$, w_t, r_{t+1}, and k_{t+1}.

The stability conditions for this world-economy model are very similar to those for the closed economy given in (21). As in the closed economy, we can solve the first-order conditions and household budget constraint for c_t^1 as a function of w_t, r_{t+1}, and ρ, and for $\bar{c}_t^1$ as a function of w_t, r_{t+1}, and $\bar{\rho}$. As in the autarky case, we can now obtain a first-order difference equation in k_t:

$$f(k_t) - k_t f'(k_t) - \frac{1}{2}\{c^1[f(k_t) - k_t f'(k_t), f'(k_{t+1}), \rho] + \bar{c}^1[f(k_t) - k_t f'(k_t), f'(k_{t+1}), \bar{\rho}]\} = k_{t+1}(1 + n). \quad (26)$$

This model will be locally stable if

$$\left|(c_w^1 + \bar{c}_w^1 - 2)kf''[2(1 + n) + (c_r^1 + \bar{c}_r^1)f'']^{-1}\right| < 1.^{11} \quad (27)$$

The interpretation of this stability condition is the same as in the closed-economy case. Again, the assumptions of normal goods and a nonnegative effect on saving of an increase in the interest rate do not suffice to establish stability. Note that, when tastes are identical, the model reduces to the closed economy. All international trade and international borrowing and lending are due to taste differences.

Combining the rate of time preference with finite-lived individuals in a Samuelson-Diamond type of overlapping-generations model may be the only analytical way to introduce different pure rates of time preference in models with international trade in goods and financial claims. If an economy consists of infinite-lived individuals who maximize the sum of the infinite stream of utility (or consumption) discounted by their rate of time preference, then in the steady state the interest rate equals the pure rate of time preference because consumption does not vary over time. Therefore, if the world consists of two such economies with different rates of time preferences, there cannot be a steady state with international mobility of financial capital. Differences in the rates of time preferences imply differences in steady-state interest rates, as in Stiglitz (1970), whose model indeed does not have financial capital mobility, while financial capital mobility implies an identical steady-state interest rate. The use of the overlapping-generations model solves this problem since, even in the

[11] Using the more general utility functions $u(c_t^1, c_t^2)$ for the home country and $\bar{u}(\bar{c}_t^1, \bar{c}_t^2)$ for the foreign country, we can solve for c_t^1 and $\bar{c}_t^1$ as functions of w_t and r_{t+1}. The stability analysis carries through unchanged for this more general case.

steady state, first-period consumption need not equal second-period consumption (even though each generation's lifetime consumption profile is the same), and thus the interest rate need not equal the rate of time preference.[12]

The Balance-of-Payments Accounts

In a two-country world, it suffices to analyze the balance-of-payments accounts of one of the countries. Without loss of generality, I focus on the home country:

$$B_t \equiv -\tilde{B}_t, \tag{28a}$$

and

$$G_t \equiv -\tilde{G}_t. \tag{28b}$$

As there is no official settlements balance, the current account surplus (deficit) of a country is identically equal to its capital account deficit (surplus).

The balance-of-trade surplus is the excess of domestic product over domestic absorption. Domestic absorption is the sum of consumption and domestic capital formation: $B_t = L_t f(k_t) - c_t^1 L_t - c_t^2 L_{t-1} - (K_{t+1} - K_t)$, or

$$b_t = f(k_t) - c_t^1 - \frac{c_{t-1}^2}{1+n} - [(1+n)k_{t+1} - k_t]. \tag{29}$$

The current account surplus is the excess of national product over domestic absorption. National product equals domestic product plus net foreign investment income. If A_t is the wealth of the home country at the beginning of period t, net claims on the rest of the world are $A_t - K_t$ and net foreign investment income $r_t(A_t - K_t)$. The country's wealth at the beginning of period t consists of the accumulated saving of the members of generation $t - 1$: $A_t = (w_{t-1} - c_{t-1}^1)L_{t-1}$, or

$$a_t = \frac{w_{t-1} - c_{t-1}^1}{1+n}. \tag{30}$$

National wealth can, and with different pure rates of time preference will, differ from the value of the domestic capital stock because of the scope for international borrowing and lending. There is a presumption that a nation consisting of people with high rates of time preference will tend to be a net foreign borrower. The truth of this presumption will be investigated in what follows.

[12] I am indebted to an anonymous referee for making this point.

The current account surplus of the home country is given by $G_t = L_t f(k_t) + r_t(A_t - K_t) - c_t^1 L_t - c_{t-1}^2 L_{t-1} - (K_{t+1} - K_t)$, or

$$g_t = f(k_t) + r_t(a_t - k_t) - c_t^1 - \frac{c_{t-1}^2}{1+n} - [(1+n)k_{t+1} - k_t]. \quad (31)$$

With constant returns to scale in production and marginal productivity factor pricing, $f(k_t) = w_t + r_t k_t$. Equation (31) can therefore be rewritten as:

$$g_t = w_t + r_t a_t - c_t^1 - \frac{c_{t-1}^2}{1+n} - [(1+n)k_{t+1} - k_t]. \quad (31')$$

Equivalently, the current account surplus can be viewed as the net foreign investment by the home country. Net foreign investment is the excess of domestic wealth accumulation (saving) over domestic capital formation, that is, $G_t = A_{t+1} - A_t - (K_{t+1} - K_t)$, or

$$g_t = (1+n)a_{t+1} - a_t - [(1+n)k_{t+1} - k_t].^{13} \quad (32)$$

Steady-State Equilibrium

The steady-state equilibrium of the two-country model is given by (11′) and (12) for the home country and their counterparts for the foreign country, the factor market equilibrium conditions (13) and (14), the stationary world capital market equilibrium conditions (33), the definition of home country net worth (34), and the balance-of-payments equations (35), (35′), (36), and (36′):

$$w - \frac{1}{2}(c^1 + \bar{c}^1) = k(1+n), \quad (33)$$

$$a = \frac{w - c^1}{1+n}, \quad (34)$$

$$b = f(k) - c^1 - \frac{c^2}{1+n} - nk, \quad (35)$$

or

$$b = (n - r)(a - k), \quad (35')$$

$$g = f(k) + r(a - k) - c^1 - \frac{c^2}{1+n} - nk, \quad (36)$$

$$g = n(a - k). \quad (36')$$

Neither the steady-state trade balance nor the steady-state current ac-

[13] $A_{t+1} = (w_t - c_t^1)L_t$; $a_{t+1} = A_{t+1}/L_{t+1}$; $c_{t-1}^2 L_{t-1} = A_t(1 + r_t)$ or $c_{t-1}^2/(1+n) = a_t(1 + r_t)$. $\therefore w_t + r_t a_t - c_t^1 - c_{t-1}^2(1+n)^{-1} = (1+n)a_{t+1} - a_t$.

count balance need be in equilibrium. As regards the current account balance, this is the result of a nonzero rate of population growth and different tastes. If, for example, the home country is a net lender to the rest of the world ($a > k$), steady-state equilibrium requires a current account surplus sufficient to maintain the real per capita value of net claims on the rest of the world, that is, a surplus of $n(a - k)$ (eq. [36']). A steady-state current account surplus will be associated with a steady-state trade balance surplus if the interest rate is less than the rate of population growth, with a steady-state trade balance deficit otherwise (35').

Because the real wage and the interest rate are equalized throughout the world economy, the present discounted value of lifetime resources and lifetime consumption is equal for residents of both countries. If and only if the two countries have identical lifetime consumption paths ($c^1 = \tilde{c}^1$ and $c^2 = \tilde{c}^2$) will there be no net foreign lending or borrowing. Each economic agent "lives within his means" in the sense that the present discounted value of lifetime resources (period 1 labor income) equals the present discounted value of lifetime consumption. Still, except when tastes are identical or population is stationary, one country will be steadily accumulating claims on the other.

We next establish the following three propositions about the steady-state equilibrium of the world economy.

Proposition 6.—The country whose residents have a higher value of consumption when young will run a steady-state current account deficit.

Proof.—From (34) we know that $a = [(w - c^1)/(1 + n)]$ and $\tilde{a} = [(w - \tilde{c}^1)/(1 + n)]$. Thus $c^1 > \tilde{c}^1$ implies $a < \tilde{a}$. Rewrite (33) as $a - k + \tilde{a} - k = 0$. Therefore $a - k = -(\tilde{a} - k) < 0$. The current account surplus is $n(a - k) < 0$. This proposition, unlike the following two, does not depend on the additively separable specification of the utility function.

Proposition 7.—The country with the higher pure rate of time preference has a steady-state current account deficit.

Proof.—From proposition 6 we know that it suffices to show that $\rho > \tilde{\rho}$ implies $c^1 > \tilde{c}^1$. This was shown in proposition 1.

Proposition 8.—The common steady-state open-economy capital-labor ratio lies between the two autarky capital-labor ratios if the model is stable under autarky and if the conditions of proposition 5 are satisfied.

Proof.—Let k^* be the common open-economy capital-labor ratio, k_H the autarky domestic capital-labor ratio, and k_F the autarky foreign capital-labor ratio. We again assume that $\rho > \tilde{\rho}$. For brevity we write steady-state consumption in period 1 as $c^1(k, \rho)$ and $\tilde{c}^1(k, \tilde{\rho})$, with $c^1_\rho, \tilde{c}^1_\rho > 0$. From the fact that consumption in period 1 is an increasing

function of ρ, it follows that $\bar{c}^1(k, \bar{\rho}) < \frac{1}{2}[c^1(k, \rho) + \bar{c}^1(k, \bar{\rho})] < c^1(k, \rho)$. To establish the truth of proposition 8 we must rank the k's determined by the following three capital market equilibrium conditions: $f(k_H) - k_H f'(k_H) - c^1(k_H, \rho) = k_H(1 + n)$; $f(k^*) - k^* f'(k^*) - \frac{1}{2}\{[c^1(k^*, \rho) + \bar{c}^1(k^*, \bar{\rho})]\} = k^*(1 + n)$; and $f(k_F) - k_F f'(k_F) - \bar{c}^1(k_F, \rho) = k_F(1 + n)$ Assume that k^* is below the lower of the two autarky capital-labor ratios, k_H. In the autarky case, at a value of k below k_H such as k^*, there would be excess demand (i.e., excess saving) in the capital market—if the model is stable. A fortiori there would be excess demand if, in the home country capital market equilibrium condition, $c^1(k^*, \rho)$ is replaced by the smaller world-average level of consumption $\frac{1}{2}[c^1(k^*, \rho) + \bar{c}^1(k^*, \bar{\rho})]$. Thus capital market equilibrium requires $k^* > k_H$. By exactly analogous reasoning, it is established that $k^* < k_F$.

Diagrammatically, the stationary equilibrium of the two-country model can be represented as in figure 3. Here OF is the stationary decentralized consumption possibility locus for the hypothetical average inhabitant of this two-country world economy, that is, someone with a lifetime consumption pattern $\frac{1}{2}(c^1 + \bar{c}^1)$, $\frac{1}{2}(c^2 + \bar{c}^2)$. It incorporates the world capital market equilibrium condition (33) and the hypothetical average world citizen's budget constraint, the sum of the

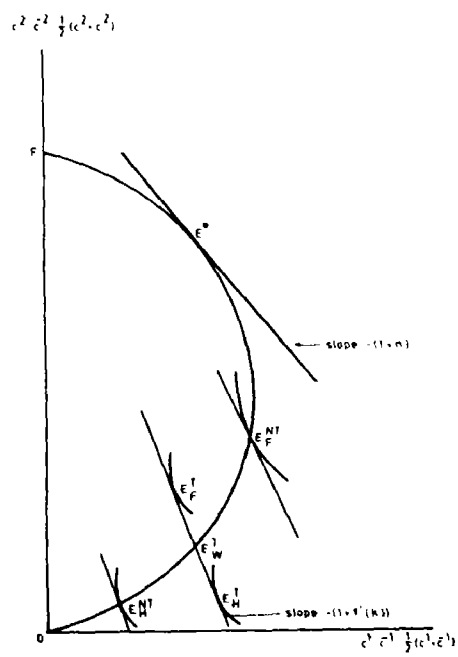


FIG. 3.—Autarky and open-economy equilibria for both countries

two countries' private budget constraints. That is, OF represents $\frac{1}{2}(c^1 + \tilde{c}^1)$ as a function of $\frac{1}{2}(c^2 + \tilde{c}^2)$, solved from:

$$f(k) - kf'(k) - \frac{1}{2}(c^1 + \tilde{c}^1) = k(1 + n), \tag{37}$$

$$f(k) - kf'(k) - \frac{1}{2}(c^1 + \tilde{c}^1) = \frac{\frac{1}{2}(c^2 + \tilde{c}^2)}{1 + f'(k)}. \tag{38}$$

The OF also represents the stationary autarky consumption possibility locus for each country individually.

While (37) and (38) must be satisfied in a full stationary equilibrium of the world economy, they are not a complete characterization of the stationary decentralized consumption possibility locus. Residents of each country must satisfy their budget constraints individually and not merely on average, across both countries as in (38). In figure 3, the indifference curve of each country is tangent to the common world capital market line. The home country, after integration into the world economy, consumes at E_H^T. Its high level of c^1 is a reflection of its high pure rate of time preference. The E_F^T is the open-economy consumption equilibrium of the foreign country with the low pure rate of time preference. Average world consumption (and world production in both countries) is at E_W^T. Consistency requires that the distance $E_F^T - E_W^T = E_W^T - E_H^T$. The home country borrows from the foreign country. From (34) and (33) we obtain that $a - k = \frac{1}{2}[(\tilde{c}^1 - c^1)/(1 + n)] < 0$. The home country's debt to the foreign country grows at the natural rate of growth. As we have drawn it, E_H^T is at a capital-labor ratio below the golden rule, that is, $r > n$. The home country therefore runs a trade balance surplus. If the equilibrium had been beyond E^*, the home country would still run a current account deficit, but with $n > r$ it would now also run a trade balance deficit. The autarky equilibrium for the home country is at E_H^{NT}, at a capital-labor ratio below the common posttrade capital-labor ratio. The autarky equilibrium for the foreign country is at E_F^{NT}, at a capital-labor ratio above the common open-economy capital-labor ratio.

Corollary.—The steady-state real wage of the capital-exporting country is below its autarky real wage and conversely for the capital-importing country. This confirms the analysis of MacDougall (1960), Kemp (1962), and Ruffin (1979).

Non-Steady-State Behavior

The behavior of the two-country model outside the steady state is not much more complicated than its steady-state behavior. There is only one state variable—the capital-labor ratio. It is easily seen from equa-

tion (39) below that the current account and trade account surpluses in period t can also be expressed as a function of k_{t-1}. The current account surplus of the home country is given by:

$$g_t = f(k_t) + r_t\left(\frac{w_{t-1} - c^1_{t-1}}{1+n} - k_t\right) - c^1_t - \frac{c^2_{t-1}}{1+n} - [(1+n)k_{t+1} - k_t]. \tag{39}$$

Will the country with the higher pure rate of time preference run a current account deficit throughout the non-steady-state adjustment process, the way it does in the steady state? Corresponding to (39) is the current account surplus equation for the foreign country:

$$\tilde{g}_t = f(k_t) + r_t\left(\frac{w_{t-1} - \tilde{c}^1_{t-1}}{1+n} - k_t\right) - \tilde{c}^1_t - \frac{\tilde{c}^2_{t-1}}{1+n} - [(1+n)k_{t+1} - k_t]. \tag{40}$$

In this two-country world $g_t + \tilde{g}_t \equiv 0$. Comparing (39) and (40), we see that with $\rho > \tilde{\rho}$, $\tilde{c}^1_t < c^1_t$, $\tilde{c}^1_{t-1} < c^1_{t-1}$, and $\tilde{c}^2_{t-1} > c^2_{t-1}$. The excess of period 1 consumption in the home country over period 1 consumption in the foreign country makes for a current account deficit in the home country. The excess of period 2 consumption in the foreign country over period 2 consumption in the home country has the opposite effect. The home country will run a current account deficit if and only if

$$\frac{r_t(c^1_{t-1} - \tilde{c}^1_{t-1})}{1+n} + c^1_t - \tilde{c}^1_t > \frac{\tilde{c}^2_{t-1} - c^2_{t-1}}{1+n}. \tag{41}$$

When the private sector budget constraints are used, this can be rewritten as

$$c^1_t - \tilde{c}^1_t > \frac{c^1_{t-1} - \tilde{c}^1_{t-1}}{1+n}. \tag{41'}$$

With $n > 0$ this condition is certainly satisfied in the steady state with $c^1_t = c^1_{t-1}$ and $\tilde{c}^1_t = \tilde{c}^1_{t-1}$ because $c^1 > \tilde{c}^1$. Outside the steady state, it will be satisfied if the increase in total consumption by the younger generation between period $t-1$ and period t in the home country exceeds that in the foreign country. This can be seen by rewriting (41′) as

$$L_t c^1_t - L_{t-1}c^1_{t-1} > L_t\tilde{c}^1_t - L_{t-1}\tilde{c}^1_{t-1}. \tag{41''}$$

The current account deficit equals the excess of domestic capital formation over national saving. This holds for both countries. With perfect international capital mobility, k_t is the same in both countries

in each period. Domestic capital formation will therefore be the same in the two countries. The home country will therefore run a current account deficit if and only if domestic saving falls short of foreign saving, that is, if and only if $L_t(w_t - c_t^1) - L_{t-1}(w_{t-1} - c_{t-1}^1) < L_t(w_t - \bar{c}_t^1) - L_{t-1}(w_{t-1} - \bar{c}_{t-1}^1)$. This is equivalent to (41″). While, with $\rho > \bar{\rho}$, each generation in the home country saves less than the corresponding generation in the foreign country, national saving in the home country need not be less in any given period than national saving in the foreign country. National saving is the sum of saving by the young and dissaving by the old. In the home country the young save less, but as a consequence the old dissave less.

IV. Gains from Participation in the World Economy

The positive analysis of the determinants of international trade and lending can be complemented in a natural way by the analysis of the gains from participation in the international economy. When the gains or losses from participation in the world economy are evaluated, welfare in the world economy will be compared with the welfare that would have been achieved under continued autarky. Both short-run effects and long-run effects are considered.

Short-Run Effects

Short-run effects are the effects on the welfare of the two generations alive during the period that the two national capital markets and goods markets are integrated. The two countries evolve under autarky up to and including period $t - 1$. International integration occurs in period t. I shall consider in detail only the case in which preexisting capital goods cannot be shifted abroad. In this case, integration of capital and goods markets in period t does not permit the owners of capital that already exists at the beginning of period t (the savings in period $t - 1$ of young members of generation $t - 1$) to move it from country to country. This physical capital was fixed in place in $t - 1$ to be used in the country of its original location. Because the physical capital stocks in the two countries in period t are predetermined, both the interest rate faced by the old generation ($t - 1$) and the wage rate faced by the young generation are predetermined and unaffected by the international integration. In this case it does not matter whether the shift toward openness is anticipated or unanticipated. The welfare of the old generation in both countries is therefore the same with openness as it would have been under continued autarky. The welfare of the younger generation will be higher (lower) under openness than under autarky if the interest rate in period t,

r_{t+1}, is higher (lower) under openness than under autarky. In figure 1, a higher interest rate pivots the budget constraint upward around w_t, while a lower interest rate pivots it downward. The quantities $\{k_\tau^H\}$, $\{k_\tau^F\}$, and $\{k_\tau\}$ are the sequences of the capital-labor ratios for the home country under autarky, for the foreign country under autarky, and for the world economy under openness, respectively. The quantities $\{r_\tau^H\}$, $\{r_\tau^F\}$, and $\{r_\tau\}$ are the associated interest rate sequences. It is assumed that at the time of integration, t, the capital-labor ratio of the home country, k_t^H, is not above that of the foreign country, k_t^F.[14]

Proposition 9.—If $k_t^H \leq k_t^F$, if consumption in period 2 is a normal good, and if an increase in the interest rate increases saving, then $k_{t+1}^H < k_{t+1} < k_{t+1}^F$ or $r_{t+1}^F < r_{t+1} < r_{t+1}^H$.

Proof.—The proof is by contradiction. Assume that $r_{t+1} > r_{t+1}^H$. From propositions 2 and 3 we know that $r_{t+1}^H > r_{t+1}^F$. Therefore, if $c_\tau^1 < 0$ and $\tilde{c}_\tau^1 < 0$, saving by generation t under openness in both the home country and the foreign country is higher than it would have been under autarky. Thus the per capita open-economy saving by generation t in the home country exceeds $(1 + n)k_{t+1}^H$, and the per capita open-economy saving by generation t in the foreign country exceeds $(1 + n)k_{t+1}^F$. With populations of the same size in both countries, average per capita saving by generation t in the world economy therefore certainly exceeds $(1 + n)k_{t+1}^H$ ($r_{t+1}^H > r_{t+1}^F$ implies $k_{t+1}^H < k_{t+1}^F$). Average per capita saving by generation t in the world economy equals the average per capita capital stock in the world economy in period $t + 1$. Therefore $k_{t+1} > k_{t+1}^H$. This contradicts the assumption that $r_{t+1} > r_{t+1}^H$. Exactly analogous reasoning yields the result that $k_{t+1} < k_{t+1}^F$.

The effect of participation in the world economy on the welfare of the two generations alive during the period in which integration occurs is therefore unambiguous. The welfare of the older generation in both countries is unaffected. The welfare of the younger generation in the high-time-preference country (the home country) is lowered, while the welfare of the younger generation in the low-time-preference country (the foreign country) is raised. During period t capital flows from the low-time-preference, low interest rate country to the high-time-preference, high interest rate country, expanding the consumption possibility set of the younger generation in the low-time-preference country and shrinking that in the high-time-preference country. Since in the high-time-preference country the older generation neither gains nor loses from integration and the

[14] This will always be the case either if the two countries started off from the same initial condition (proposition 4) or if, from arbitrary initial conditions, the system had been evolving under autarky for a sufficiently long period of time (propositions 2, 3, and 5).

younger generation loses, a move toward integration will always be vetoed by the residents of the high-time-preference country if they vote according to their preferences—unless an appropriate transfer from the low-time-preference country is forthcoming.[15]

Long-Run Effects

It is not possible, except in special cases, to make unambiguous welfare comparisons by comparing steady-state utility levels—the ranking of the stationary utility levels may not be the same as the ranking of the utility levels achieved during the transition from one steady state to another. Nevertheless, the stationary utility levels achieved under autarky and openness are of interest. They are analyzed next. There are several cases to consider, illustrated in figures 4–9.

Case 1.—The autarky capital-labor ratios of both countries are above the golden-rule capital-labor ratio. In figures 4 and 5, the autarky equilibrium of the home country is at E_H^{NT}, that of the foreign country at E_F^{NT} (proposition 5). The golden-rule capital-labor ratio is at E^*. The open-economy consumption equilibrium of the home country is at E_H^T, that of the foreign country at E_F^T. World production equilibrium is at E_W^T, at a capital-labor ratio between the two closed-economy capital-labor ratios. The foreign country (the low-time-preference country) has a higher stationary level of utility under openness than under autarky (figs. 4 and 5). Two factors contribute to this. First, the open-economy capital-labor ratio, while still in the inefficient region, is lower than under autarky. The extent of overaccumulation of capital is reduced, and the total amount of resources available for consumption is increased. Second, the foreign country is a net lender at an interest rate higher than the autarky interest rate. The home country (the high-time-preference country) can have either a lower (fig. 4) or a higher (fig. 5) stationary level of utility under openness. The degree of overaccumulation is increased by international integration. Against that, the home country is a net borrower at an interest rate below the autarky interest rate.

[15] If preexisting capital goods can be shifted abroad, capital goods accumulated during period $t - 1$ out of the saving of young members of generation $t - 1$ can be shifted between countries so as to establish a common interest rate, r_t, and marginal product of capital $f'(k_t) = r_t$ for period t. It now makes a difference whether the move toward openness in period t was anticipated in period $t - 1$. If perfect foresight prevails and if $k_{t-1}^H \leq k_{t-1}^F$, it is easily shown that the old generation alive during the transition to openness (generation $t - 1$) will be worse off under openness than under continued autarky in the home country, better off in the foreign country. The effect on the welfare of the young generation (t) is ambiguous. It is therefore no longer necessarily the case, as with nonshiftable capital, that the high-time-preference home country will always veto a move toward openness.

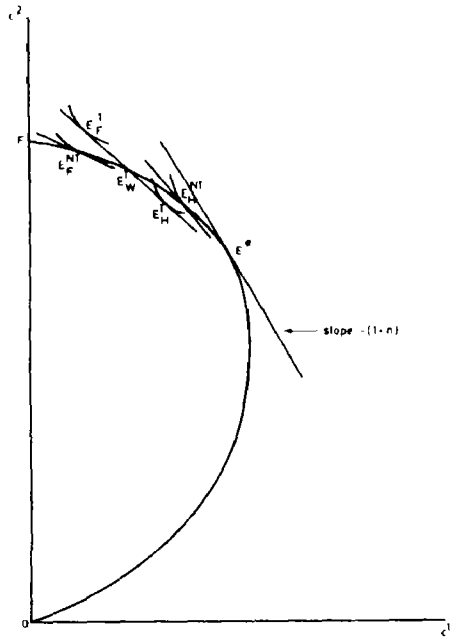


FIG. 4.—Long-run welfare comparison: both autarky capital-labor ratios above the golden rule.

Case 2.—The autarky capital-labor ratios of both countries are below the golden-rule capital-labor ratio. This case is illustrated in figures 6 and 7. The high-time-preference country (the home country) has a higher stationary utility level under openness than under autarky. Under openness, the capital-labor is higher but remains below the golden-rule level. The total amount of resources available for consumption is therefore increased. The country also is a net borrower at an interest rate lower than the autarky interest rate. The low-time-preference country (the foreign country) can have either a lower (fig. 6) or a higher (fig. 7) stationary level of utility under openness. Its capital-labor ratio is reduced in the efficient region, lowering the total amount of consumable resources. Against that, the foreign country is a net lender at an interest rate higher than its autarky interest rate.

Case 3.—The autarky capital-labor ratios are on opposite sides of the golden rule. Two subcases need to be distinguished.

Case 3a.—The open-economy capital-labor ratio is above the golden-rule capital-labor ratio. In this case the analysis for the low-time-preference country (the foreign country) is the same as in case 1: Its stationary utility level is unambiguously higher under openness

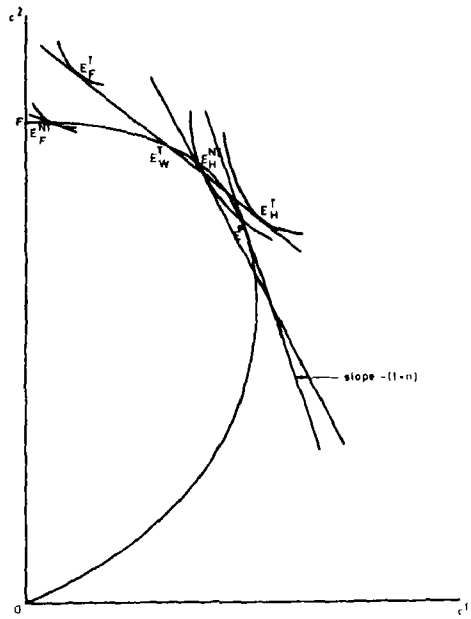


FIG. 5.—Long-run welfare comparison: both autarky capital-labor ratios above the golden rule.

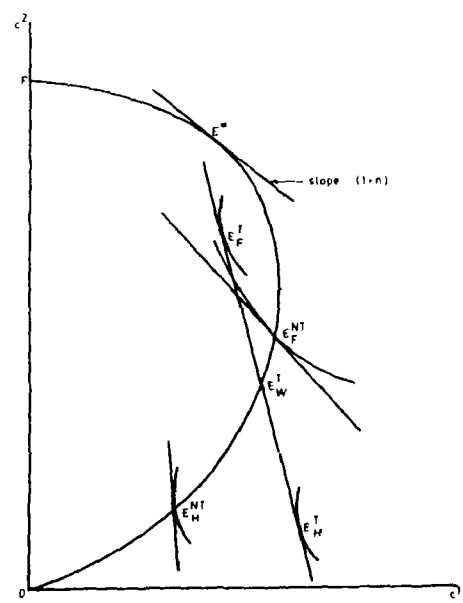


FIG. 6.—Long-run welfare comparison: both autarky capital-labor ratios below the golden rule.

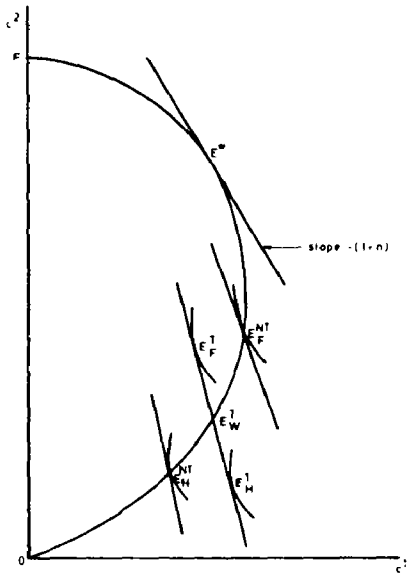


FIG. 7.—Long-run welfare comparison: both autarky capital-labor ratios below the golden rule.

than under autarky. The stationary utility level of the high-time-preference country (the home country) can be either lower (fig. 8) or higher (fig. 9) under openness. There is a stronger presumption toward a higher utility level in this case than in case 1, when the autarky capital-labor ratio of the high-time-preference country was above the golden-rule level. In case 3*a* the open-economy capital-labor ratio at E_W^T, while in the inefficient region, may still correspond to a stationary volume of consumable resources above that corresponding to E_H^{NT}, which is below the golden rule. If it does, E_H^T will certainly correspond to a higher stationary utility level than E_H^{NT} because in addition the high-time-preference country is a net borrower at an interest rate below the autarky interest rate.

Case 3b.—The open-economy capital-labor ratio is below the golden-rule capital-labor ratio. The analysis for the high-time-preference country (the home country) is now the same as in case 2: Its stationary utility level is unambiguously higher under openness than under autarky. The low-time-preference country's stationary utility level can be either higher or lower under openness. No diagrams are provided as they can easily be constructed by analogy with figures 8 and 9.

An important clue to these welfare comparisons is provided by the

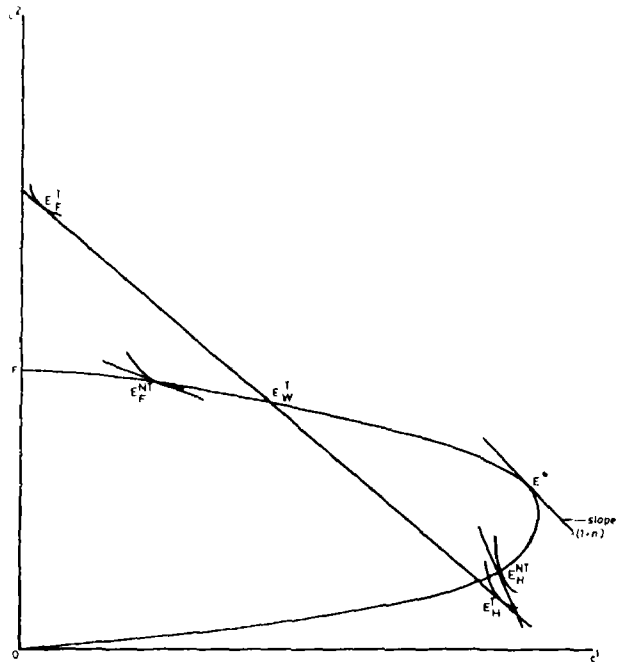


FIG. 8.—Long-run welfare comparison: the autarky capital-labor ratios straddle the golden rule.

fact that international trade and financial mobility do not expand the consumption possibility set of the "average" world consumer under openness beyond that available to the individual countries under autarky. The *OF* in figures 3 and 4 is the stationary, decentralized consumption possibility locus for each of the countries under autarky and for the integrated world economy. The same applies in the short run: There are no potential gains in production from specialization. The only difference made by openness is that a worldwide capital market is substituted for two national capital markets. The equilibrium in this worldwide capital market will be an average of the two autarky equilibria in the national capital markets.

V. Conclusion

Diamond's overlapping-generations model is extended to a two-country world. A number of results are obtained. The country whose residents consume more in the first period of their lives (at a given wage rate and interest rate) has a steady-state current account deficit if the rate of population growth is positive. The country with the

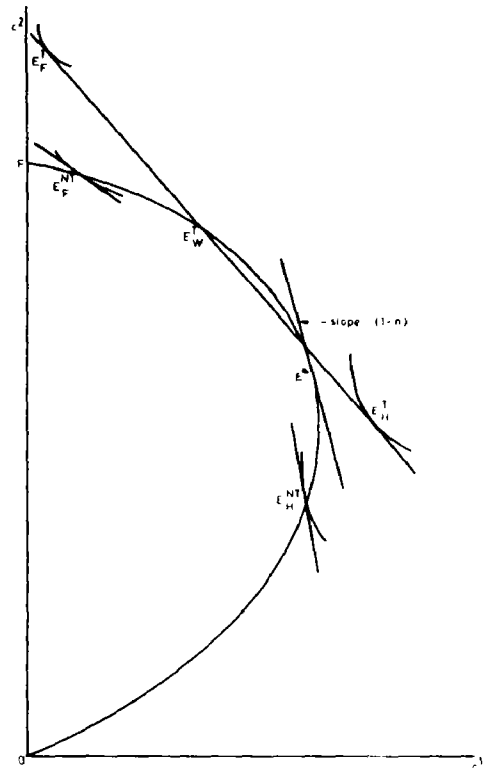


FIG. 9.—Long-run welfare comparison. the autarky capital-labor ratios straddle the golden rule.

higher pure rate of time preference has a steady-state current account deficit. The common stationary open-economy capital-labor ratio lies between the two autarky capital-labor ratios, with the lower autarky capital-labor ratio associated with the country whose residents have the higher pure rate of time preference. Outside the steady state, the country with the higher pure rate of time preference will not necessarily be the one to run a current account deficit. If preexisting capital is not shiftable between countries, the welfare of the old generation alive when the two autarkic countries join in an open economy is unaffected by this international economic integration. The welfare of the young generation in the country with the high pure rate of time preference is lower under openness than under autarky, while in the country with the low pure rate of time preference it is higher.

The comparison of stationary utility levels under autarky and openness yields two unambiguous results. If both the autarky and the

open-economy capital-labor ratios of the low-time-preference country are above the golden-rule level, its stationary utility level will be higher under openness. If both the autarky and the open-economy capital-labor ratios of the high-time-preference country are below the golden-rule level, its stationary utility level will be higher under openness. If the first of these conditions holds, the move toward openness can be said to constitute a potential Pareto improvement for the low-time-preference country, because it always gains in the short run. Even if the second condition holds, openness cannot be said to be a Pareto-superior regime for the high-time-preference country, as it always incurs a welfare loss in the short run.

The open-economy version of the overlapping-generations model can be extended in a number of directions. Within the context of the one-commodity model, budgetary and financial policy can be studied; this includes public sector lending and borrowing, taxation of wage and property income, social security, etc. The two-commodity structure of the Heckscher-Ohlin model can be grafted onto the overlapping-generations framework.[16] This allows potential gains from trade through specialization in production. If one of the goods is a capital good, the distinction between international mobility of financial capital—international lending and borrowing—and international trade in capital goods becomes important. One could have either or both. The two-commodity structure permits the analysis of tariffs, quotas, and other trade policies, familiar from the static Heckscher-Ohlin model, in addition to the financial and budgetary policies that can be studied in the one-commodity overlapping-generations model. Finally, uncertainty can be introduced into the model—most interestingly into the two-commodity version. Without futures markets for commodities, households have to make labor supply and saving decisions while young, without knowing the future spot prices of commodities with certainty. The scope for further developments appears to be considerable.

References

Borts, George H. "A Theory of Long-Run International Capital Movements." *J.P.E.* 72, no. 4 (August 1964): 341–59.

Buiter, Willem H. "Government Finance in an Overlapping Generations Model with Gifts and Bequests." In *Social Security versus Private Saving*, edited by George M. von Furstenberg. Cambridge, Mass.: Ballinger, 1979.

Carmichael, Jeffrey. "The Role of Government Financial Policy in Economic Growth." Ph.D. thesis, Princeton Univ., 1979.

[16] In a paper by Fried (1980), an overlapping-generations model with two tradable, nondurable commodities is analyzed. This model does not contain a variable capital stock.

Cass, David, and Yaari, Menahem E. "A Re-examination of the Pure Consumption Loans Model." *J.P.E.* 74, no. 4 (August 1966): 353–67.

Diamond, Peter A. "National Debt in a Neoclassical Growth Model." *A.E.R.* 55 (December 1965): 1126–50.

Findlay, Ronald. "An 'Austrian' Model of International Trade and Interest Rate Equalization." *J.P.E.* 86, no. 6 (December 1978): 989–1007.

Fischer, Stanley, and Frenkel, Jacob A. "Investment, the Two-Sector Model and Trade in Debt and Capital Goods." *J. Internat. Econ.* 2 (August 1972): 211–33.

Fried, Joel. "The Intergenerational Distribution of the Gains from Technical Change and from International Trade." *Canadian J. Econ.* 13 (February 1980): 65–81.

Hori, Hajime, and Stein, Jerome L. "International Growth with Free Trade in Equities and Goods." *Internat. Econ. Rev.* 19 (February 1977): 83–100.

Ihori, Toshihiro. "The Golden Rule and the Role of Government in a Life Cycle Growth Model." *A.E.R.* 68 (June 1978): 389–96.

Kemp, Murray C. "Foreign Investment and the National Advantage." *Econ. Rec.* 38 (March 1962): 56–62.

MacDougall, G. D. A. "The Benefits and Costs of Private Investment from Abroad: A Theoretical Approach." *Econ. Rec.* 36 (March 1960): 13–35.

Neher, Philip A. "International Capital Movements along Balanced Growth Paths." *Econ. Rec.* 46 (September 1970): 393–401.

Onitsuka, Yusuke. "International Capital Movements and the Patterns of Economic Growth." *A.E.R.* 64 (March 1974): 24–36.

Ruffin, Roy J. "Growth and the Long-Run Theory of International Capital Movements." *A.E.R.* 69 (December 1979): 832–42.

Samuelson, Paul A. "An Exact Consumption Loan Model of Interest with or without the Social Contrivance of Money." *J.P.E.* 66, no. 6 (December 1958): 467–82.

Stiglitz, Joseph E. "Factor Price Equalization in a Dynamic Economy." *J.P.E.* 78, no. 3 (May/June 1970): 456–88.

Mobility Costs, Frictional Unemployment, and Efficiency

Peter A. Diamond

Massachusetts Institute of Technology

With imperfect job information flows, it is plausible that the distribution of job offerings becomes more attractive when there are more vacancies and more unemployed. With word-of-mouth communication, this condition is derived. Given this condition, steady-state equilibrium is not efficient, with welfare increased by the introduction of unemployment compensation even though all agents are risk neutral. In this way workers become more selective in the jobs they accept.

Moving and training costs play a significant role in job-taking decisions.[1] Even if workers were equally productive in all jobs, these costs would make it sometimes worthwhile for an unemployed worker to refuse a job offer while waiting for a more attractive offer. The rate at which workers are offered jobs with different moving costs depends on the decisions of other workers as to which jobs to refuse. This externality implies that equilibrium will not generally be efficient. With the plausible assumption that job offerings become more attractive on average when the number of available jobs increases, efficiency increases when workers are induced to pass up jobs with relatively high moving costs (by unemployment compensation, e.g.).

These results are derived in a model of steady-state search equilibrium similar to one that has been used elsewhere (see Diamond and Maskin 1979 and in press; Diamond 1980*b*, 1980*c*; Mortensen, in press).

Comments from Joel Yellin and a referee, research assistance by Drew Fudenberg, and financial support from NSF are gratefully acknowledged.

[1] For a discussion of the role of employment in mobility decisions, see Bartel (1979).

[*Journal of Political Economy*, 1981, vol. 89, no. 4]
© 1981 by The University of Chicago 0022-3808/81/8904-0014$01.50

Earlier analyses have focused on decisions which affect the rate at which workers receive job offers, taking as given the distribution of the quality of job offers. Here the arrival rate of job offers is taken as given, with the quality distribution endogenous. It is assumed that there is a fixed-coefficients technology. Workers and firms are taken to be risk neutral and to face a common exogenous interest rate. Both job termination and the arrival of new job offers are modeled as Poisson processes with fixed parameters. It is assumed that the wage and sharing of moving costs are negotiated,[2] with workers and firms equally good bargainers. The steady-state unemployment level is then determined by the job-taking decisions of workers. More stringent standards for job taking raise the vacancy rate, which in turn improves the distribution of job offers. Thus there can be multiple steady-state equilibria. From any steady-state equilibrium, inducing a permanent further decrease in the moving costs which workers are willing to bear raises the present discounted value of output in the economy. The optimal unemployment compensation benefit is derived in Section VI. Presentation of these results takes as given the relationship between the unemployment rate and the quality of job offers. In Section VII, a simple model of job-information flows is presented to derive an example of the way higher job availability might improve average job-offer quality. An example with an exponential distribution of moving costs is examined in detail in Sections VIII and IX. The example results in optimal benefits in the neighborhood of 60 percent of the wage.

I. Employment

Assume that all jobs are the same, with a fixed-coefficients technology that results in a flow of output, y. All jobs are subject to a risk of termination at the constant breakup rate b. This represents exogenous factors such as a transfer of the worker or the job opportunity to a different location. We ignore the possibility of the worker and the job moving together. We also ignore endogenous reasons for job termination, such as quitting for a better job or laying off workers to hire different workers.

At initial employment, the wage is negotiated, as is the sharing of setup costs. For simplicity we shall assume complete symmetry between workers and jobs (including equal numbers) implying that the net gain from commencing production is shared equally between

[2] Negotiation is assumed to be instantaneous, ruling out one route by which the presence of other workers affects a job seeker—that of waiting while a decision is made about employing a different worker.

worker and employer. In the absence of labor disutility, capital user cost, and unemployment compensation, this implies equal sharing of the flow of output and so a wage equal to $y/2$. Workers are assumed to be risk neutral and to face a constant interest rate r. With these assumptions, it is appropriate to focus attention on the expected present discounted value of earnings less moving costs. Denote by W_E and W_U the expected discounted value of earnings less moving costs for employed and unemployed workers, respectively. With complete symmetry assumptions about both finding and losing jobs, with an infinite expected life, and with analysis restricted to the steady state, W_E and W_U will not vary over time or workers.

For an employed worker, the rate of interest times expected earnings equals the wage less the expected capital loss from job termination:[3]

$$rW_E = y/2 - b(W_E - W_U). \tag{1}$$

The next step in the analysis is to consider job taking and so the determination of W_U.

II. Job Taking

Assume that unemployed workers learn about job opportunities with an exogenous[4] arrival rate, a. We assume no search costs and no decisions of an individual which affect this arrival rate. While the productivity of all jobs is taken to be the same for all workers, jobs differ across workers in the setup costs before production begins.[5] These costs reflect moving costs when workers must relocate to take new jobs and specific training costs before production can begin. The pattern of a fixed cost followed by constant output can be viewed as an approximation to the increased output that comes with on-the-job learning. We ignore the variation among jobs in commuting costs since these represent a variation in flow benefits rather than setup costs.

Denote by $G(c,u)$ the distribution of setup costs associated with jobs an individual learns about when the unemployment rate is u. A higher unemployment rate (and higher vacancy rate) is assumed to improve the distribution of moving costs in the sense that $G_u(c,u) > 0$ where

[3] Considering a worker as an asset, eq. (1) is stated as the rate of return times the value of an asset equals the flow return plus expected capital gain. It can be derived alternatively from explicit calculation of the present value of expected returns as in Diamond (1980*b*) or as the limit of a discrete time process: $W_E = (1 + r\Delta t)^{-1}[y\Delta t/2 + b\Delta t W_U + (1 - b\Delta t)W_E]$.

[4] Below, we will allow a to vary with the unemployment rate.

[5] For a detailed analysis of individual choice with setup costs, see Loikkanen and Pursiheimo (1979).

$G(c,u)$ is positive and less than one.[6] In Sections VII and VIII we derive this assumption from specific models of information flows. We take the distribution to be constant over time for a given worker. With symmetry between workers and jobs, setup costs are assumed to be equally divided between worker and job. If an unemployed worker accepts any job with a setup cost less than c^*, then, in the absence of unemployment benefits, we can write the expected discounted value of net earnings implicitly as

$$rW_U = a\int_0^{c^*} (W_E - W_U - c/2)dG(c,u). \tag{2}$$

That is, the interest rate times the expected value of earnings equals the expected gain from job taking, less setup costs.[7]

The choice problem for the individual worker is the selection of c^* to maximize W_U. Naturally, this involves accepting any job for which the setup cost is less than the expected gain from job taking:[8]

$$c^* = 2(W_E - W_U) = \left(y + a\int_0^{c^*} cdG\right)/[r + b + aG(c^*,u)], \tag{3}$$

where (1) and (2) have been solved to give the implicit equation for c^*. Workers are more willing to bear setup costs when output is greater, job finding is more difficult, the interest rate is lower, expected job duration (b^{-1}) is longer, or the unemployment rate is lower. That is, from implicit differentiation of (3) we have $\partial c^*/\partial y > 0$, $\partial c^*/\partial a < 0$, $\partial c^*/\partial r < 0$, $\partial c^*/\partial b < 0$, and $\partial c^*/\partial u < 0$.

III. Equilibrium

In steady-state equilibrium, the aggregate rate of job finding must equal the rate of job losing. Denoting the unemployment rate by u, this gives us

$$b(1 - u) = auG(c^*,u). \tag{4}$$

That is, the job breakup rate times the proportion employed equals the job acceptance rate times the proportion unemployed. Since G_u is assumed positive, we have c^* decreasing with u. One would not generally expect to find the rate, a, at which workers learn of potential jobs

[6] To interpret this assumption one needs to compare alternative economies with different steady-state employment and vacancy rates. It is not appropriate to consider a single economy over a business cycle, since in that case a rise in unemployment is accompanied by a decline in vacancies rather than a move in the same direction.

[7] As a discrete time process, we would have $W_U = (1 + r\Delta t)^{-1}(a\Delta t\{G(c^*,u)W_E + [1 - G(c^*,u)]W_U - \int_0^{c^*} c/2dG\} + (1 - a\Delta t)W_U)$.

[8] Eq. (3) comes from joint wealth maximization of firm and worker and does not depend on the equal sharing rule.

to be independent of the unemployment and vacancy rates (which are equal by assumption). We would expect a to increase with u across steady states. Since the implications of this relationship have been explored elsewhere, we take a to be exogenous here but note the implied differences in footnotes. Similar considerations hold for b.

Since c^* decreases with u in both (3) and (4), we have the possibility of multiple steady-state equilibria, that is, multiple solutions to (3) and (4). When more jobs are available (higher u), anticipated mobility costs are lower (higher G) and individuals are more selective in the jobs they take (lower c^*). Greater selectivity by workers, in turn, raises the unemployment rate. This is shown in figure 1, where equations (3) and (4) are drawn.

We note that an individual will not accept a job with an expected return below setup costs. Thus the chosen cutoff c^* is less than $y/(r + b)$. Even if expected future setup costs are close to zero, the potential loss in forgone wages implies that the chosen cutoff c^* will never be less than $y/(r + b + a)$. If all offered jobs are accepted, the equilibrium unemployment rate is $b/(a + b)$. If none is accepted, the unemployment rate goes to one.

IV. Steady-State Output

All steady-state equilibria lie along the curve given in (4). At any point on this curve we can calculate the steady-state output level per person. There are $1 - u$ employed per person giving a gross output flow per person of $(1 - u)y$. Moving costs per person equal the rate of new job taking per person of auG times the average cost of moving, $\int_0^{c^*} cdG/G$. Thus net output, Q, satisfies

$$Q = (1 - u)y - au\int_0^{c^*} cdG. \tag{5}$$

Differentiating with respect to u, with c^* given as an implicit function of u by (4), evaluating at an equilibrium where (3) holds, and using

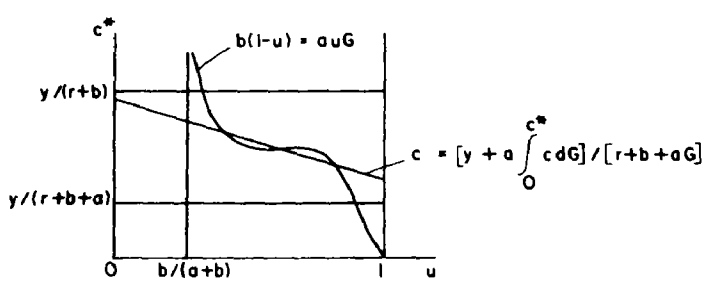


FIG. 1

integration by parts, we have

$$\frac{dQ}{du} = -y - a\int_0^{c^*} c\,dG - auc^*g\frac{dc^*}{du} - au\int_0^{c^*} cg_u dc$$
$$= -rc^* + au\int_0^{c^*} G_u dc, \tag{6}$$

where $g = G_c$. The second term reflects the externality in lower moving costs of a higher unemployment rate. The first factor reflects the absence of discounting in steady-state comparisons.

V. Dynamics

The economy analyzed here cannot move directly from one steady state to another. While it is interesting to compare alternative steady states which might have occurred, proper policy analysis requires consideration of the comparative statics of the actual path of the economy. We shall analyze the effects of a policy which controls c^* directly. That is, we consider an economy where job acceptance behavior is sufficiently closely monitored to make it a government control variable. Below we will examine how unemployment compensation can be used to induce the same steady-state equilibrium without monitoring job acceptance behavior. These two modes of control correspond to stylized versions of the German and American economies.

Unemployment grows by job terminations and declines by job acceptances:

$$\dot{u} = b(1-u) - auG(c^*,u). \tag{7}$$

Starting at a steady-state equilibrium given by (3) and (4) we shall calculate the change in the present discounted value of aggregate net output, W, from a permanent differential change in c^*, with unemployment given by (7).

That is, we want to calculate the derivative with respect to c^* of

$$W \equiv \int_0^\infty e^{-rt}\left\{[1-u(t)]y - au(t)\int_0^{c^*} c\,dG[c;\,u(t)]\right\}dt, \tag{8}$$

where $u(t)$ satisfies (7) with an initial condition satisfying (3) and (4). Calculating this derivative (see Diamond 1980*a*), we have

$$\frac{\partial W}{\partial c^*} = \frac{-auc^*g}{r} + \left(\frac{y + a\int_0^{c^*} c\,dG + au\int_0^{c^*} cg_u dc}{r}\right)\left(\frac{aug}{r + b + aG + auG_u}\right). \tag{9}$$

Using (3) we can write this as[9]

$$\frac{\partial W}{\partial c^*} = \left(\frac{aug}{r}\right)\left[-c^* + \frac{(r + b + aG)c^* + au\int_0^{c^*} cg_u dc}{r + b + aG + auG_u}\right]$$

$$= \frac{-(a^2u^2g)\left(c^*G_u - \int_0^{c^*} cg_u dc\right)}{r(r + b + aG + auG_u)} \tag{10}$$

$$= \frac{-a^2u^2g}{r}\left(\frac{\int_0^{c^*} G_u dc}{r + b + aG + auG_u}\right) < 0.$$

Thus a government policy to move the economy out of a steady-state equilibrium in the direction of higher unemployment raises efficiency by improving the average quality of job offers.[10]

VI. Unemployment Compensation

There are several modes of intervention which will raise the equilibrium unemployment rate and so the efficiency of this economy. By increasing the importance of adjustment costs relative to the financial gains of employment, the government can induce greater selectivity in job acceptance. Thus, taxing output without allowing a deduction for adjustment costs or subsidizing unemployment or vacancies will have the desired end. Of these we shall analyze unemployment compensation. In addition to making workers more selective in accepting jobs, unemployment compensation raises the wage by one-half the unemployment compensation benefit given the negotiation assumptions we have made here.

From the perspective of the worker and employer, the net gain from production is $y - B$, where B is the unemployment benefit. With this net gain shared equally, the firm receives $0.5(y - B)$ and the worker receives a payment for forgoing unemployment compensation of B in addition to half the net gain, $0.5(y - B)$. Rewriting the two value equations (1) and (2), we have[11]

[9] If a is a function of u, (10) becomes

$$\frac{\partial W}{\partial c^*} = \frac{-au^2g}{r}\left[\frac{\int_0^{c^*}(aG_u + a'G)dc}{r + b + aG + auG_u + a'uG}\right].$$

[10] It is important to remember that we are considering policies that increase both vacancies and unemployed together. For an analysis of the gain from aggregate demand stimulation, see Diamond (1980*c*).

[11] To derive (11) formally, we need first to reconsider the equal-sharing rule above. That was derived from equations for W_E and W_U, similar equations for the values of

$$rW_E = 0.5(y + B) - b(W_E - W_U)$$
$$rW_U = B + a\int_0^{c^*} (W_E - W_U - c/2)dG(c,u). \tag{11}$$

Then, the chosen cutoff cost satisfies

$$c^* = 2(W_E - W_U) = \left(y - B + a\int_0^{c^*} cdG\right)/(r + b + aG). \tag{12}$$

Increasing the unemployment benefit by \$1.00 decreases the cutoff level of costs selected at a constant unemployment rate by $(r + b + aG)^{-1}$ dollars. The induced increase in unemployment magnifies this effect.

Setting equation (9) equal to zero, we can derive the cutoff cost level which would be optimal in a steady state. That is, an optimal trajectory will, asymptotically, have the cutoff cost level, c^*, satisfying[12]

$$c^* = \frac{y + a\int_0^{c^*} cdG + au\int_0^{c^*} cg_u dc}{r + b + aG + auG_u}. \tag{13}$$

Solving (12) and (13), we can derive the unemployment compensation benefit which holds asymptotically on the efficient trajectory (i.e., which supports the asymptotic optimum):

$$B^* = \frac{\left(y + a\int_0^{c^*} cdG\right)auG_u - au\int_0^{c^*} cg_u dc(r + b + aG)}{r + b + aG + auG_u}. \tag{14}$$

Substituting from the first-order condition for c^*, B^* can be written alternatively as[13]

$$B^* = au\left(c^*G_u - \int_0^{c^*} cg_u dc\right) = au\int_0^{c^*} G_u dc > 0. \tag{15}$$

Passing up a job offer alters the trajectory of the economy and so generates a pattern of externalities that varies over time. The present

filled and vacant jobs, W_F and W_V, and an equal sharing of surplus value, $W_E - W_U = W_F - W_V$. With unemployment compensation and equal sharing of setup costs the four value equations can be expressed in terms of the wage, w:

$$rW_E = w - b(W_F - W_U),$$
$$rW_F = y - w - b(W_F - W_V),$$
$$rW_U = B + a\int (W_E - W_U - c/2)dG,$$
$$rW_V = a\int (W_F - W_V - c/2)dG.$$

Combining these with the equal gain rule gives (11).

[12] From (13) and the first line of (6), we note that $dQ/du < 0$ at u^*.

[13] If a varies with u, (15) becomes $B^* = u\int_0^{c^*} (aG_u + a'G)dc$.

discounted value of the social gain from accepting a job with setup cost c can be derived by differentiating W, in (8), with respect to u. Evaluated at a steady state this gives

$$-\frac{\partial W}{\partial u} - c = \frac{y}{r + b + aG + auG_u} - c. \tag{16}$$

In an efficient equilibrium this social gain must equal the private gain $2(W_E - W_U) - c$. Equating the social and private gains, using (12), we have an alternative derivation of the asymptotically optimal unemployment compensation.

VII. Discrete Example

Having examined the implications of the dependence of the distribution of adjustment costs of job offers on the unemployment rate, we now consider an example of how word-of-mouth communication of job availability can generate such dependence.[14] In this section we consider an example with two locations. Specific training costs are c_1. There are no moving costs if the job and worker are in the same location. There are moving costs c_2 if the job and worker are in different locations. For the next section we consider an example where c is distributed exponentially with a parameter that depends linearly on the unemployment rate.

Assume that employed workers learn of a job opportunity in the same location and inform their unemployed friends of its existence. Assume that the process of job communication is such that attempted communication about any vacancy is a Poisson process with constant parameter a. (In practice it is likely that there are more attempted communications of new job vacancies than of old ones.) Each potential communicator knows n workers who could fill this job. Assume that each of the n has a probability u of being unemployed and an independent probability p of being in the same location. With probability $(1 - u)^n$ all n friends are employed and there is no one to tell of the job. If the communicator does have unemployed friends, he only tells a friend in the other location if none of his unemployed friends are in the same location. That is, he tells the friend to whom the information is most valuable. The probability of his telling someone in the other location is $[(1 - up)^n - (1 - u)^n]$—the probability of no unemployed friends with the same location less the probability of no unemployed friends. Then $1 - (1 - up)^n$ is the probability of telling a

[14] For previous analyses of word-of-mouth communication, see Boorman (1975) and Satterthwaite (1979).

friend in the same location. The greater the unemployment rate, the greater the probability of telling a friend in the same location.

With everyone following the same behavior rule, there are two candidates for equilibrium: accept only jobs at the same location, or accept any job offer. Let u_1 and u_2 be the unemployment rates under these two behavior rules. Then the equilibria satisfy

$$\begin{aligned} b(1 - u_1) &= au_1[1 - (1 - u_1 p)^n] \\ b(1 - u_2) &= au_2[1 - (1 - u_2)^n]. \end{aligned} \tag{17}$$

For some parameter values both of these equilibria will exist. Rather than pursue this example in more detail, we turn to a similar example with a continuum of locations, which can be considered to be around a circle. It is then assumed that the likelihood of knowing an individual is exponentially distributed with the distance to his location and that communication goes to the unemployed person for whom the moving costs are smallest.

VIII. Exponential Example: Steady-State Properties

We assume that c has an exponential distribution with coefficient nu. This example is chosen to fit the discussion above since the minimum of a random sample of size nu from the exponential distribution with coefficient 1 is exponential with coefficient nu. Thus we assume that

$$G(c,u) = 1 - e^{-nuc}. \tag{18}$$

With this distribution, the two equilibrium equations (3) and (4) satisfy

$$c^*(r + b + a) = y + \frac{a}{nu}(1 - e^{-nuc^*}) \tag{19}$$

$$b(1 - u) = au(1 - e^{-nuc^*}). \tag{20}$$

There exists a unique solution to this pair of equations with u between $b/(a + b)$ and one and c^* between $y/(r + b + a)$ and $y/(r + b)$.

In a steady state (whether equilibrium or not), the net output flow satisfies, from equation (5),

$$\begin{aligned} Q &= (1 - u)y - an^{-1}[1 - e^{-nuc^*}(1 + nuc^*)] \\ &= (1 - u)y - \frac{b(1 - u)}{nu} \\ &\quad - \left(\frac{a}{n}\right)\left[1 - \frac{b(1 - u)}{au}\right] \ln\left[1 - \frac{b(1 - u)}{au}\right]. \end{aligned} \tag{21}$$

The asymptotically optimal unemployment compensation satisfies, from equation (15),

$$B^* = \left(\frac{a}{nu}\right)[1 - e^{-nuc^*}(nuc^* + 1)]$$
$$= \frac{b(1-u)}{nu^2} + \left(\frac{a}{nu}\right)\left[1 - \frac{b(1-u)}{au}\right] \ln\left[1 - \frac{b(1-u)}{au}\right]. \tag{22}$$

From (21) and (22) we note that

$$Q = (1 - u)y - uB^*. \tag{23}$$

Alternatively this follows from (15) since, with the exponential distribution, $\int uG_u = \int cg$. From (13) we note that at the steady-state optimum the maximal acceptable moving costs satisfy

$$\frac{y}{c^*} = \frac{b}{u^*} + r. \tag{24}$$

To find the optimal unemployment rate, we solve (20) and (24) simultaneously.

Some examples are shown in table 1. For these calculations, parameters were chosen for a, b, and r. Next an equilibrium unemployment rate in the absence of government intervention, u, was chosen. This implied a particular value for the product ny, and the remaining calculations were done for this value. In addition to steady-state comparisons, the last column of table 1 reports the percentage change in the present discounted value of net output along the trajectory from the equilibrium steady state to the optimal one where the fraction of accepted jobs, G, is held constant at its asymptotically optimal value. As detailed in the next section, the increase in net output along the trajectory exceeds the difference between steady states.

As a guide to interpreting the table, consider the fourth row. In an economy where the unemployed receive 10 job offers a year, the expected duration of a job is 5 years, the interest rate is 5 percent, and the equilibrium unemployment rate is 4 percent, 3 percent of gross output is spent on moving costs, workers are just willing to spend 34 percent of a year's output on moving, and workers accept 48 percent of jobs they hear about. The asymptotically optimal unemployment rate is 4.7 percent, which can be induced by unemployment compensation equal to 60 percent of the wage. In this equilibrium workers accept 41 percent of job offers representing an 18 percent increase in the expected duration of unemployment, and steady-state net output is higher than in the no-compensation equilibrium by 0.23 percent

TABLE 1

EXPONENTIAL EXAMPLE

a	b	r	u	$\frac{Q}{(1-u)y}$	$\frac{c^*}{y}$	$G(c^*,u)$	u^*	$\frac{B^*}{w}$	G^*	$\frac{Q^*}{Q}$	$\frac{W'}{W}$
10	.2	.05	.01965	.996	.12	.998	.01971	.31	.994	1.0000	1.0000
...	...	...	.02	.99	.13	.98	.0205	.42	.96	1.0001	1.0001
...	...	...	.03	.98	.25	.65	.0344	.59	.56	1.0013	1.0014
...	...	...	.04	.97	.34	.48	.0470	.60	.41	1.0023	1.0024
...	...	...	.05	.96	.43	.38	.0593	.61	.32	1.0032	1.0034
...	...	...	.06	.95	.52	.31	.0715	.61	.26	1.0040	1.0043
...	...	...	.07	.94	.61	.27	.0835	.60	.22	1.0048	1.0051
...	...	...	.08	.93	.69	.23	.0954	.60	.19	1.0054	1.0059
...	...	...	.09	.93	.77	.20	.1071	.59	.17	1.0061	1.0066
...	...	...	.10	.92	.85	.18	.1188	.59	.15	1.0066	1.0073
5	.2	.05	.04	.99	.27	.96	.0415	.45	.92	1.0002	1.0002
10	...	...	..	.97	.34	.48	.0470	.60	.41	1.0023	1.0024
20	...	...	...	.97	.36	.24	.0484	.62	.20	1.0031	1.0032
50	...	...	...	.96	.37	.10	.0491	.63	.08	1.0036	1.0037
100	...	...	...	.96	.37	.05	.0493	.64	.04	1.0037	1.0038
1,000	...	...	...	.96	.38	.005	.0494	.64	.004	1.0038	1.0039
10	.05	.05	.04	.97	1.40	.12	.0487	.61	.10	1.0028	1.0032
..	.1	...	..	.97	.71	.24	.0483	.62	.20	1.0029	1.0031
...	.2	...	...	97	.34	.48	.0470	.60	.41	1.0023	1.0024
10	.2	.02	.04	.97	.35	.48	.0471	.61	.40	1.0024	1.0025

with a gain of 0.24 percent along the constant G trajectory from the initial equilibrium to the asymptotic optimum.

The examples show a surprisingly consistent pattern. When the equilibrium unemployment rate is very close to $b/(a + b)$, the minimum achievable,[15] the optimal unemployment compensation is small. As u rises, B^*/w rises very rapidly, reaching the neighborhood of 60 percent when u is about one percentage point above the achievable minimum. The ratio B^*/w stays in the neighborhood of 60 percent for all calculated values, which included values of u as integer percentages up to 10 percent. The same pattern arose for all calculated values of a, b, and r.

IX. Continuous Example Dynamics

The transition from the equilibrium steady state to the optimal one shows a larger change in net output than does the steady-state comparison. To see this, let us consider a policy of changing the proportion of jobs accepted by an unemployed worker. (We shall also consider the optimal policy below.) If workers are accepting all jobs with costs below $c^*(t)$, then the proportion of jobs accepted equals $1 - e^{-nu(t)c^*(t)}$. Let us consider the policy which holds this proportion constant over time at the level that occurs in the asymptotic steady state which we write as G^*. Then the economy follows the differential equations

$$\begin{aligned} \dot{u}(t) &= b[1 - u(t)] - au(t)G^* \\ \frac{\dot{c}^*(t)}{c^*(t)} &= \frac{-\dot{u}(t)}{u(t)}. \end{aligned} \tag{25}$$

We assume an initial condition at time t_0 of the steady-state equilibrium. The immediate effect of the change in policy is to decrease aggregate setup costs without changing gross output. Over time the unemployment rate rises, decreasing gross output but keeping aggregate setup costs constant. This pattern is shown in figure 2. To calculate W along this trajectory, begin by considering aggregate setup costs. These equal

$$au(t)\int_0^{c^*(t)} cnu(t)e^{-nu(t)c}dc = \frac{a}{n}\{1 - e^{-nu(t)c^*(t)}[nu(t)c^*(t) + 1]\}. \tag{26}$$

Since $u(t)c^*(t)$ is constant, we have

$$Q(t) = Q^* + y[u^* - u(t)]. \tag{27}$$

[15] When $a = 10$ and $b = 0.2$, $b/(a + b) = .019608$.

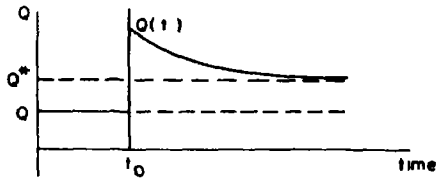


FIG. 2

Solving (25) for $u(t)$, we have

$$u(t) = u^* + (u - u^*)e^{-(bt/u^*)}. \tag{28}$$

Thus the present discounted value of output satisfies

$$W' = \int_0^\infty e^{-rt}Q(t)dt = \frac{Q^*}{r} + \frac{y(u^* - u)}{\left(r + \frac{b}{u^*}\right)}. \tag{29}$$

Since the optimal unemployment rate exceeds the equilibrium rate, the value of output along this path exceeds its value at the optimal steady state. The excess of W' over equilibrium output, W, is shown in table 1. The addition to the value of unemployment compensation from analyzing the dynamic path is small but noticeable for the examples calculated.

In addition to considering this path, which was chosen for its ease of analysis, it is interesting to analyze the optimal path assuming the government could control $c^*(t)$. The choice problem is

$$\max_{c^*(t)} \int_0^\infty e^{-rt}\left([1 - u(t)]y - \frac{a}{n}\{1 - e^{-nu(t)c^*(t)}[1 + nu(t)c^*(t)]\}\right) dt$$
$$\text{s.t. } \dot{u}(t) = [1 - u(t)]b - au(t)[1 - e^{-nu(t)c^*(t)}]. \tag{30}$$

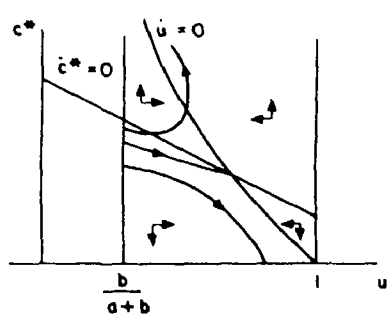


FIG. 3

The first-order condition for $\dot{c}^*$ is

$$\dot{c}^*(t) = (r + b)c^*(t) - y + ac^*(t)[1 - e^{-nu(t)c^*(t)}]. \tag{31}$$

The phase diagram for the optimal path is shown in figure 3. Along the optimal path, $c^*(t)$ is falling while $u(t)$ is rising. Their product, and so the proportion of job offers accepted, is rising. In terms of figure 2 the optimal path of net output converges to Q^* more rapidly than the path with uc^* constant, starting from a higher initial net output level.

Baily (1977) and Flemming (1978) have analyzed optimal unemployment compensation assuming risk-averse workers and no externalities in the labor allocation process. This paper takes the opposite tack of assuming risk-neutral workers and externalities. It seems likely that workers are risk averse and externalities are present, making a much stronger case for unemployment compensation.

References

Baily, Martin N. "Unemployment Insurance as Insurance for Workers." *Indus. and Labor Relations Rev.* 30, no. 4 (July 1977): 495–504.

Bartel, Ann P. "The Migration Decision: What Role Does Job Mobility Play?" *A.E.R.* 69, no. 5 (December 1979): 775–86.

Boorman, Scott A. "A Combinatorial Optimization Model for Transmission of Job Information through Contact Networks." *Bell J. Econ.* 6, no. 1 (Spring 1975): 216–49.

Diamond, Peter A. "An Alternative to Steady State Comparisons." *Econ. Letters* 5 (1980): 7–9. (*a*)

———. "Wage Determination in Search Equilibrium." Working Paper no. 253, M.I.T., Dept. Econ., January 1980. (*b*)

———. "Aggregate Demand Management in Search Equilibrium." Working Paper no. 268, M.I.T., Dept. Econ., November 1980. (*c*)

Diamond, Peter A., and Maskin, Eric. "An Equilibrium Analysis of Search and Breach of Contract. I. Steady States." *Bell J. Econ.* 10 (Spring 1979): 282–316.

———. "An Equilibrium Analysis of Search and Breach of Contract. II. A Nonsteady State Example." *J. Econ. Theory*, in press.

Flemming, John S. "Aspects of Optimal Unemployment Insurance: Search, Leisure, Savings and Capital Market Imperfections." *J. Public Econ.* 10, no. 3 (December 1978): 403–25.

Loikkanen, Heikki A., and Pursiheimo, Ulla. "On an Extended Model of Job Search." Working Paper no. 118, Univ. Helsinki, Dept. Econ., 1979.

Mortensen, Dale T. "The Matching Process as a Noncooperative Bargaining Game." In *Economics of Information and Uncertainty*, edited by John J. McCall. Chicago: Univ. Chicago Press (for Nat. Bur. Econ. Res.), in press.

Satterthwaite, Mark. "Consumer Information, Equilibrium Industry Price, and the Number of Sellers." *Bell J. Econ.* 10, no. 2 (Autumn 1979): 483–502.

Comments

International Transmission of a Real Shock under Flexible Exchange Rates: A Comment

Betty C. Daniel

Tulane University

In a recent article in this journal, Hamada and Sakurai (1978) claim to have demonstrated that a recession in a foreign country, generated either by a reduction in monetary growth or by a decline in productivity, is transmitted to the domestic country as a temporary stagflation. They emphasize the wage-price spiral as the transmission mechanism which operates under fixed and flexible exchange rates. In particular, any fall in foreign output causes an increase in the relative price of foreign goods. This puts upward pressure on home wages through a Phillips curve, which includes expected inflation of consumer prices; employment and output fall while inflation increases. The article deserves acclaim for focusing on the neglected "wage-price spiral" transmission mechanism; however, it is misleading in its assertion that a reduction in output abroad, whatever its cause, has temporary effects at home.

The purpose of this note is to demonstrate that the wage-price spiral transmission mechanism, following a decline in foreign productivity, generates a permanent fall in domestic output as well as temporary stagflation. The productivity decline, which permanently lowers foreign output, *permanently* deteriorates the domestic terms of trade and *permanently* reduces domestic output by reducing the purchasing power of labor's wage. The deterioration in the terms of trade leaves the real producer wage (wage divided by the price of the domestic output) unaffected. However, the higher price of imports

I wish to thank my students Rebecca Morton and Vincent Salvo for drawing attention to the inconsistency implied by *temporary* domestic effects of a *permanent* foreign real shock. I also wish to thank Koichi Hamada for excellent comments on an earlier draft. Partial financial support from the Murphy Institute of Political Economy is gratefully acknowledged.

[*Journal of Political Economy*, 1981, vol. 89, no. 4]
© 1981 by The University of Chicago. 0022-3808/81/8904-0009$01.50

reduces labor's real wage (wage divided by the consumer price index), creating an excess demand for labor. The nominal wage is bid up, and employment and output fall.[1]

The source of the authors' result is an incompletely specified model. Their model is adequate for monetary shocks (their primary interest) but not for real shocks. Extension of their model to meaningfully incorporate real shocks requires a respecification of their Phillips curve.

The authors specify the domestic and foreign Phillips curves by letting the rate of wage inflation in each country ($\hat{w}$, $\hat{w}^*$, where an asterisk refers to a foreign variable) equal expected inflation in the consumer price index (π, π^*) plus a function (ψ) of output produced in each country (X_1, X_2^*, where subscripts 1 and 2 refer to the commodities produced in the domestic and foreign countries, respectively). They are expressed as

$$\hat{w} = \psi(X_1) + \pi, \tag{1-1}$$

$$\hat{w}^* = \psi(X_2^*) + \pi^*. \tag{1-2}$$

(Equation numbers correspond to those of the authors unless preceded by the letter A. The latter are alternative equations proposed by myself.)

The long-run requirement that $\hat{w} = \pi$ implies that there is only one value of X_1 consistent with long-run equilibrium. The analogous relation holds for the foreign country. This implication is clearly inconsistent with allowing real shocks to have permanent long-run output effects. In particular, it is inconsistent with the author's claim that a productivity decline in the foreign country permanently reduces X_2^*. This problem is easily solved by respecifying the Phillips curve to allow wage inflation to depend on the deviation of output from long-run equilibrium as well as on expected inflation according to

$$\hat{w} = \psi(X_1 - \bar{X}_1) + \pi, \tag{A1-1}$$

$$\hat{w}^* = \psi(X_2^* - \bar{X}_2^*) + \pi^*, \tag{A1-2}$$

where a bar denotes a long-run equilibrium.

It is now necessary to specify the determinants of long-run output implied by the authors' model. Steady state is characterized by equality in the rates of growth of money (θ), prices ($\hat{p}$), and wages ($\hat{w}$). Output is determined by the authors' production functions,

$$X_1 = N^{\beta}, \qquad 0 < \beta < 1; \tag{3-1}$$

$$X_2^* = N^{*\beta^*}, \qquad 0 < \beta^* < 1; \tag{3-2}$$

[1] This effect is discussed in Salop (1974), Daniel (1977), and Leiderman (1979).

where X denotes output, N denotes employment, and β is the employment elasticity of output.

Labor demand is determined to equate the nominal wage (w) with the value of marginal product according to

$$w = p_1 \beta N^{\beta-1}, \tag{4-1}$$

$$w^* = p_2^* \beta^* N^{*\beta^*-1}, \tag{4-2}$$

where p denotes price.

Prices are determined by the assumption that expenditure on each commodity is a constant fraction of total expenditure. This assumption together with the authors' assumption that expenditure equals income yields the authors' equations,

$$p_1 X_1 = vM, \tag{13-1}$$

$$p_2^* X_2^* = vM^*. \tag{13-2}$$

To close the model for long-run output, a specification of the long-run supply of labor is necessary. I assume labor supplied is a positive function of the real wage according to

$$\bar{N} = (w/p)^{\delta}, \tag{a}$$

$$\bar{N}^* = (w^*/p^*)^{\delta^*}, \tag{b}$$

where δ and $\delta^* > 0$ are the long-run real wage elasticities of the labor supply, and p and p^* are the consumer price indices in the domestic and foreign countries and are defined by the authors as $p = p_1^{\alpha} p_2^{1-\alpha}$ and $p^* = p_1^{*\alpha} p_2^{*1-\alpha}$. This specification is in keeping with the short-run Phillips curve whereby a slower growth rate of nominal wages, relative to the price level, reduces the real wage and reduces output. By allowing labor supply to care about the real wage in the long run as well as in the short run, the possibility that a disturbance which alters the real wage can permanently alter output is admitted. This is preferable to assuming away long-run effects by fixing long-run labor supply while allowing short-run labor supply to respond.

Equations (3-1), (3-2), (4-1), (4-2), (13-1), (13-2), (a), and (b) can be solved to yield the reduced-form solution for long-run output:

$$\bar{X}_1 = \gamma(\overset{+}{\beta}, \overset{+}{\beta^*}), \tag{c}$$

$$\bar{X}_2^* = \gamma^*(\overset{+}{\beta}, \overset{+}{\beta^*}). \tag{d}$$

It is important to notice that a fall in β^* permanently reduces domestic output because it deteriorates the home country's terms of trade thereby reducing labor's real wage.[2]

[2] This can be proved by totally differentiating the log linear version of the model

It is now possible to meaningfully analyze a shock to foreign productivity. The respecified Phillips curve together with explicit allowance for changes in productivity require rewriting the authors' equations (20) and (22) as

$$\hat{X}_1 = \beta\left[\theta - \psi(X_1 - \bar{X}_1) + \left(1 + \frac{\ln X_1}{\beta}\right)\hat{\beta} - \pi\right], \tag{A20}$$

$$\hat{X}_2^* = \beta^*\left[\theta^* - \psi(X_2^* - \bar{X}_2^*) + \left(1 + \frac{\ln X_2^*}{\beta^*}\right)\hat{\beta}^* - \pi^*\right], \tag{A22}$$

where ^ denotes a proportional rate of change. The long-run requirement that $\theta - \pi = \hat{X}_1 = \hat{\beta} = 0$ ($\theta^* - \pi^* = \hat{X}_2 = \hat{\beta}^* = 0$) implies that $X_1 = \bar{X}_1$ ($X_2 = \bar{X}_2$). Anything which changes the long-run level of output requires an eventual corresponding change in $\bar{X}_1$ ($\bar{X}_2$).

The model is completed with the assumption of adaptive inflationary expectations which yields the authors' equations (21) and (23). They are unchanged by the foregoing modifications:

$$\dot{\pi} = \gamma\{\theta - [\alpha\hat{X}_1 + (1 - \alpha)\hat{X}_2^*] - \pi\}, \tag{21}$$

$$\dot{\pi}^* = \gamma\{\theta^* - [\alpha\hat{X}_1 + (1 - \alpha)\hat{X}_2^*] - \pi^*\}. \tag{23}$$

Figure 1 depicts the relationships between π and X_1 necessary for $\hat{X}_1 = 0$ and $\dot{\pi} = 0$ for given β, $\hat{\beta}$, $\bar{X}_1$, $\hat{X}_2^*$, and θ and the corresponding relationships in the foreign country. Equations (A20) and (A22) imply that for $\hat{X}_1 = 0$, inflation must be negatively correlated with output. The same type of relation exists in the foreign country. The $\dot{\pi} = 0$ relation is determined by substituting equation (A20) into (21) and analogously for $\dot{\pi}^* = 0$. Dynamics are described by arrows following the authors.

Now consider a fall in foreign productivity. This can be represented

composed of eqq. (4-1), (4-2), (13-1), (13-2), (a), and (b) where the exchange rate is normalized at unity.

$$\frac{dp_2}{d\beta^*} = \frac{-[\beta(1 - \alpha) + (\eta - \beta)](1 + \eta^* \ln N^*)p_2}{\Delta} < 0.$$

$$\frac{dp_1}{d\beta^*} = \frac{-\beta(1 - \alpha)(1 + \eta^* \ln N^*)p_2}{\Delta} < 0,$$

$$\frac{dN^*}{d\beta^*} = \frac{\{(1 - \alpha)[(\beta/\beta^*) + \eta \ln N^*] + (1/\beta^*)(\eta - \beta)\}N^*}{\Delta} > 0,$$

$$\frac{dN}{d\beta^*} = \frac{(1 - \alpha)(1 + \eta^* \ln N^*)N}{\Delta} > 0,$$

where $\Delta = \alpha\beta^*(\eta - \beta) + (1 - \alpha)(\eta^* - \beta^*)\beta + (\eta - \beta)(\eta^* - \beta^*)$, $\eta = (1 + \delta)/\delta$, and $\eta^* = (1 + \delta^*)/\delta^*$. With $\beta = \beta^*$ initially, as the authors assume, it is obvious that a reduction in β^* deteriorates the domestic terms of trade and reduces foreign employment relatively more than domestic employment. Foreign output also falls relatively more from eqq (3-1) and (3-2).

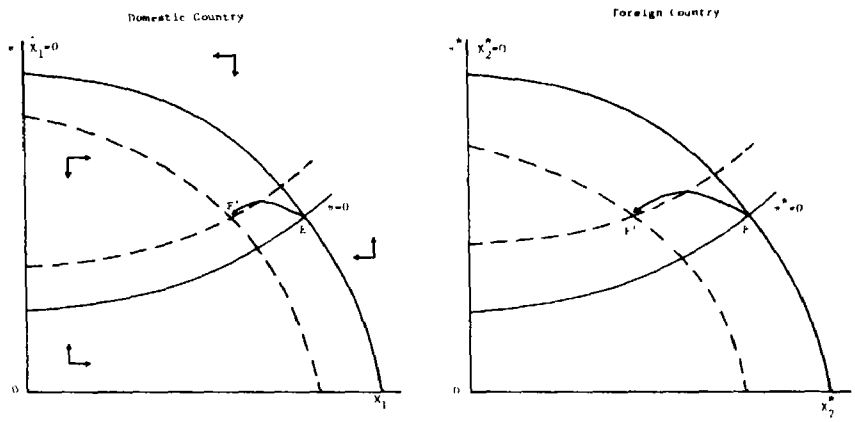


FIG. 1

by a shock to β^*. In equation (A22), $\hat{\beta}^*$ falls temporarily and $\bar{X}_2^*$ falls permanently. Both effects shift $\hat{X}_2^* = 0$ left in figure 1, and since the fall in $\bar{X}_2^*$ is permanent, it remains left of its original position. In equation (23), the fall in β^* stimulates inflationary expectations by reducing $\hat{X}_1$ (from [A20] and [c]) and by reducing $\hat{X}_2^*$ (from [A22] and [d]); $\pi^* = 0$ shifts up to the left permanently. Under the assumption that $\hat{\beta}^* \neq 0$ for a very short period of time, the foreign adjustment path can be drawn as by the authors in figure 1.

The domestic adjustment path and final equilibrium are not as described by the authors. The fall in $\bar{X}_2^*$, which occurs due to the productivity shock, causes a permanent increase in its relative price (see n. 2 above). This reduces home labor's real wage and permanently reduces $\bar{X}_1$. Therefore, $\hat{X}_1 = 0$ also shifts left. It shifts less than $\hat{X}_2^* = 0$. The permanent increase in the relative price of foreign output implies that $\bar{X}_1$ falls less than $\bar{X}_2^*$; $\dot{\pi} = 0$ also shifts left due to the decline in $\bar{X}_1$ and $\bar{X}_2^*$ caused by the fall in foreign productivity. The adjustment path and long-run equilibrium are similar to those abroad. The stagflation is temporary in both countries, but the fall in output is permanent.[3]

In conclusion, this note confirms that the wage-price spiral does cause a fall in foreign output, whatever its cause, to be transmitted as stagflation in the short run, as the authors claim. However, it should be noted that the wage-price spiral also causes a permanent foreign

[3] Hamada has noted that the fall in output could be permanent under the assumption that labor supply is independent of the real wage in the long run. This assumption seems less appropriate in a paper about the wage-price spiral transmission mechanism in which the short-run dynamics depend crucially on the effect of changes in the real wage on labor supply.

productivity deterioration to be transmitted as a long-run decline in domestic output. The permanent terms of trade deterioration, implied by the productivity shock, reduces the purchasing power of the domestic workers' wage leading to less employment and output.

References

Daniel, Betty C. "Inflation and Unemployment in Open Economics." International Finance Discussion Paper no. 114, Board Governors Federal Reserve System, December 1977.

Hamada, Koichi, and Sakurai, Makato. "International Transmission of Stagflation under Fixed and Flexible Exchange Rates." *J.P.E.* 86, no. 5 (October 1978): 877–95.

Leiderman, Leonardo. "Expectations and Ouput-Inflation Tradeoffs in a Fixed-Exchange-Rate Economy." *J.P.E.* 87, no. 6 (December 1979): 1285–1306.

Salop, Joanne. "Devaluation and the Balance of Trade under Flexible Wages." In *Trade, Stability, and Macroeconomics: Essays in Honor of Lloyd A. Metzler*, edited by George Horwich and Paul A. Samuelson. New York: Academic Press, 1974.

Market Constraints as a Rationale for the Friedman-Savage Utility Function

Elie Appelbaum

University of Western Ontario

Eliakim Katz

Bar-Ilan University

I. Introduction

Since the classic paper by Friedman and Savage (1948), it has generally been accepted that the observed fact that individuals or firms[1] participate in unfair lotteries and other forms of unfair risk taking[2] may be explained by a section in the individual's utility function in which the individual shows risk preference rather than risk aversion. In their model, Friedman and Savage specify a utility function which is, in turn, concave, convex, and concave, thus allowing for simultaneous purchase of insurance (risk aversion) and participation in lotteries (risk preference).

However, while their specified utility function is indeed capable of explaining observed behavior, it does seem to be rather unsatisfactory in that it is an ad hoc specification. It is the purpose of this paper to suggest a set of circumstances which give rise to a Friedman-Savage–type utility function.

In particular, it is shown that when certain capital market imperfections exist the utility function defined over intermediate wealth should be distinguished from and may have different properties than the one defined on final wealth. Then, even if we accept the common assumption made in the literature that individuals are risk averse, that is, that their utility function is concave over final wealth, it is still possible that they participate in unfair gambling (and, of course, may also purchase insurance).

An earlier version of this paper was presented at the European Meeting of the Econometric Society, Athens, September 1979. We wish to thank the editor and an anonymous referee for their helpful comments

[1] The word "individual" is, throughout this paper, used to denote economic units, including both individuals and firms.

[2] The term "unfair lottery" in the paper encompasses all investment opportunities offering actuarially unfair risks.

[*Journal of Political Economy*, 1981, vol. 89, no. 4]
© 1981 by The University of Chicago 0022-3808/81/8904-0010$01.50

The motivation behind our endeavor to provide an institutional explanation for gambling is a reluctance (shown, e.g., by Stigler and Becker [1977]) to use non-well-behaved utility functions as explanations for economic phenomena without some a priori reasons for assuming that the utility function is of an unusual shape. Clearly all phenomena in economics can be technically explained by recourse to tastes. Thus, for example, individuals will necessarily show risk preference over a section of their utility function which is assumed convex. However, is there any a priori reason to assume that there may be such a section? It is a positive answer to this question which this paper attempts to provide.

In Section II we discuss certain capital market imperfections and their implications about the nature of rates of return on investment. In Section III we show how the Friedman-Savage utility function may emerge when utility is defined on intermediate wealth and there exist capital market imperfections. The possibility of gambling with risk aversion and of simultaneous gambling and insurance therefore immediately arises.

II. Rates of Return and Capital Market Imperfections

A number of recent studies discuss the effects of uncertainty, imperfect information, and various transaction costs on capital markets and show that the result may be that capital markets are characterized by certain imperfections. These imperfections are usually in terms of the nonexistence of certain markets and the fact that there may not be free and equal access to other markets. Uncertainty and the possibility of costly default may lead lenders to introduce collateral requirements or take default costs into account in their loan rates. Consequently, it can be said that one has to have certain assets (providing collateral services) in order to have easier or cheaper access to capital markets.

Barro (1976) and Benjamin (1978) derive these conclusions explicitly and show that market imperfections will lead to loan rates being functions of loan sizes and available collateral. In particular, they derive a loan-supply function which is constant for some initial range and then becomes an increasing and convex function of loan sizes.

In the same vein, Jaffee and Modigliani (1969) show that, as a result of uncertainty and imperfect and costly information, capital markets may be characterized by credit rationing. They show that beyond a certain point loan rates will generally depend on loan sizes, with a possible upper bound on loan sizes. Furthermore, they provide empirical evidence supporting these types of imperfections.

Empirical evidence indicating capital market imperfections is also

given by Eckstein (1961) and Nerlove (1968), who find capital markets to be characterized by differential rates of return on given investments.

The important implication of the capital market imperfections is that individuals face a variety of capital market constraints and, therefore, do not have free and equal access to the market. Consequently, capital provides additional services by either weakening the accessibility constraints or by reducing the cost of the acquisition of capital. By providing collateral services, capital, therefore, increases the set of "feasible activities" or reduces their cost, so that its "full rate of return" should take this additional role into account.

A consequence of these capital market imperfections is that for some levels of asset holdings the rates of return on assets may depend on the levels of the asset holdings, even from an individual's (rather than an aggregate) point of view. The range within which this is likely to happen is when wealth levels are low, but not below some minimum level. When an individual's wealth is below some minimum level the capital market constraints may be so effective that except for the possibility of obtaining small loans the market is in effect inaccessible. Since small loans can usually be obtained at constant loan rates (as is shown in Jaffee and Modigliani [1969], Barro [1976], and Benjamin [1978]), the additional role of asset holding is ineffective within this initial range, and consequently the rates of return will not depend on asset holdings. However, as an individual's asset holdings increase above some minimum level, his higher level of wealth will provide him with the additional "collateral services," and rates of return within this range will depend on the levels of wealth.

Since rates of return are usually not known with certainty, the dependence of rates of return on asset holdings is in a probabilistic sense. In other words, the rates-of-return probability distribution function changes with asset holdings, that is, the distribution is conditional on asset holdings.[3]

In addition to the above considerations, rates of return may be increasing functions of wealth (over some initial range) for other reasons as well. First, a large number of fixed costs may be incurred in effecting profitable investment. Information costs incurred in locating high-return investments may be considerable, and a large element of these costs may be relatively fixed. Transaction costs, where again the large part is fixed, are also likely to comprise a high percentage of small investments. In the presence of these costs the rate of return will be an increasing function of the investment. Second, there may exist

[3] Thus, e.g., an increase in asset holdings will over some range make the probability distribution more stochastically dominant.

significant indivisibilities which (especially in view of the capital market constraints) will imply increasing returns to scale.

In view of these considerations, we conclude that individuals, especially those with small or moderate levels of wealth, may very often face various capital market constraints which lead to rates of return being functions of wealth. Furthermore, the relationship between wealth and the rates of return is such that rates of return are (at least over some range) increasing functions of wealth.

III. Gambling with Risk Aversion

Having discussed some of the capital market imperfections and the constraints they impose on individuals, we now consider their effects on the individual's attitudes toward risk taking.

In order to focus on the gambling problem and to separate it from the investment problem, we assume that the rate of return is nonrandom. This assumption is not crucial and does not change the results. For the sake of simplicity, we also assume that there is only one asset. A discussion of the case with many assets and random rates of return can be found in Appelbaum and Katz (1979), where the same results are derived.

In line with the discussion in the previous section, it will be assumed that the individual faces a rate of return R on his investment, A, such that

$$R = R(A) \qquad \begin{aligned} R' &= 0 \qquad 0 \leq A \leq A^* \\ R' &> 0 \quad \text{for } A^* \leq A \leq A^{**} \\ R' &= 0 \quad \text{for} \qquad A > A^{**}, \end{aligned} \tag{1}$$

where R' is the partial derivative of R. In other words, the rate of return is constant over some initial range, increasing over some subsequent range,[4] and then constant again.

An individual investing an amount A will end up with final wealth W, such that

$$W = A[1 + R(A)]. \tag{2}$$

Clearly, convexity of $R(A)$ over some range is sufficient, but not necessary, for the convexity of $W(A)$ over the same range; both convex and concave rate-of-return functions may lead to a convex $W(A)$.[5]

[4] This type of return function is discussed also in Blinder (1974) and Appelbaum and Harris (1978).

[5] It does, however, seem reasonable that at least over some initial range (when R is not constant) $R'' > 0$, i.e., the function is convex. The convexity of $R(A)$ at the initial range would, e.g., follow from the convexity of the loan-supply function derived by Jaffee and Modigliani (1969), Barro (1976), and Benjamin (1978).

Let us now consider the individual's decision problem. Following the literature,[6] we assume the individual has a utility function defined on final wealth, $U(W)$, and that

$$U' > 0, \quad U'' < 0; \tag{3}$$

that is, the individual is risk averse.

Now, consider the individual's utility function when it is defined over A—his intermediate wealth. This is given by

$$V(A) = U\{A[1 + R(A)]\} \tag{4}$$

and

$$V'(A) = U' \cdot (1 + R + AR'), \tag{5}$$

$$V''(A) = U'' \cdot (1 + R + AR')^2 + U' \cdot (2R' + AR''). \tag{6}$$

Therefore, in accordance with our above discussion, $V' > 0$ for all values of A. For $A < A^*$ and $A > A^{**}$, V'' is clearly negative. Within the range $A^* < A \leq A^{**}$ the second term on the right-hand side of (6) may be positive and, hence, V'' may be positive. If this occurs there will be a section of $V(A)$ that will be convex and, hence, V will show the Friedman-Savage shape of being in turn concave, convex, and concave.

The result may also be derived diagrammatically, as is shown in figure 1. The curve $V(A)$ drawn in the first quadrant is clearly a Friedman-Savage type of utility function which allows both gambling and insurance in intermediate wealth to take place simultaneously.

Even if $V(A)$ is concave within the range $A^* < A < A^{**}$, we can still get a convex region, as in the Friedman-Savage utility function, since $V(A)$ will, in general, have a kink at A^*. Thus, although in this case $V(A)$ is made of concave segments only, it will nevertheless not be globally concave. This case is shown in figure 2, where again it is possible to observe gambling and insurance simultaneously.

Hence, we have provided an explanation of the shape of the Friedman-Savage utility function without suggesting that an individual is anything but everywhere risk averse. Rather than focus on the tastes of the individual, we have focused on his institutional constraints.[7]

[6] See, e.g., Arrow 1970.

[7] An alternative way of obtaining these results is to employ the state preference approach and consider the effects of capital market imperfections on the set of acceptable gambles. It can be easily shown (see Appelbaum and Katz 1979) that the isoutility curves defined over two random states are not necessarily convex, and thus both gambling and insurance buying may occur simultaneously.

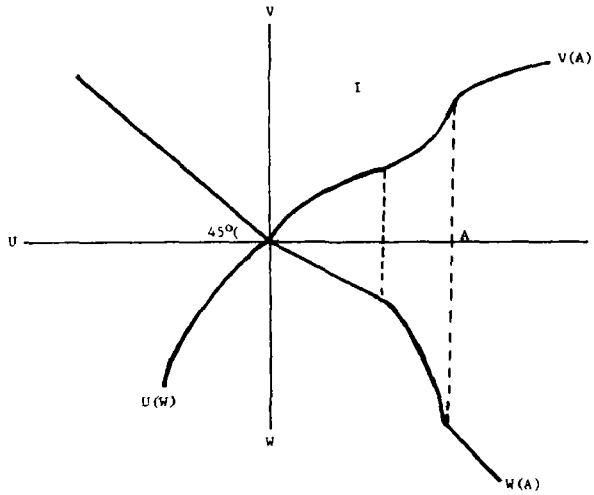


FIG. 1

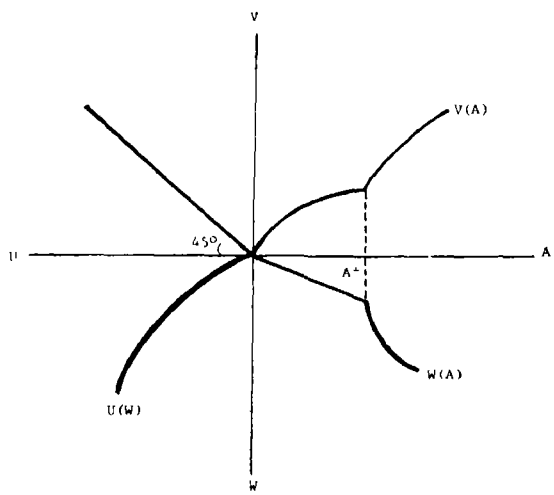


FIG. 2

IV. Conclusion

This paper provides a possible rationale for the Friedman-Savage utility function. It suggests that the existence of market imperfections, in particular in the capital markets, may impose various constraints on individuals and thus affect their behavior. The existence of these constraints could lead individuals to participate in unfair gambling, since this may reduce the implicit costs of the constraints.

Our explanation of gambling focuses on the constraints facing an individual rather than on his preferences, and consequently it does not require a modification of the standard assumption that individuals are risk averse. Of course, it may be that individuals show risk preference over a section of their utility function. Such a section is not, however, necessary in order to derive a Friedman-Savage type of utility function.

References

Appelbaum, Elie, and Harris, Richard C. "Imperfect Capital Markets and Life-Cycle Saving." *Canadian J. Econ.* 11 (May 1978): 319–24.

Appelbaum, Elie, and Katz, Eliakim. "Imperfect Capital Markets and Gambling with Risk Aversion." Discussion Paper no. 7912, Univ. Western Ontario, May 1979.

Arrow, Kenneth J. *Essays in the Theory of Risk-bearing.* Chicago: Markham, 1970.

Barro, Robert J. "The Loan Market, Collateral, and Rates of Interest." *J. Money, Credit and Banking* 8 (November 1976): 439–56

Benjamin, Daniel K. "The Use of Collateral to Enforce Debt Contracts." *Econ. Inquiry* 16 (July 1978): 333–59.

Blinder, Alan S. *Toward an Economic Theory of Income Distribution.* Cambridge, Mass.: MIT Press, 1974.

Eckstein, Otto. "A Survey of the Theory of Public Expenditure Criteria. Reply." In Conference of the Universities–National Bureau Committee for Economic Research, *Public Finances: Needs, Sources, and Utilization.* Princeton, N.J.: Princeton Univ. Press, 1961.

Friedman, Milton, and Savage, Leonard J. "The Utility Analysis of Choices Involving Risk." *J.P.E.* 56, no. 4 (August 1948): 279–304.

Jaffee, Dwight M., and Modigliani, Franco. "A Theory and Test of Credit Rationing." *A.E.R.* 59 (December 1969): 850–72.

Nerlove, Marc. "Factors Affecting Differences among Rates of Return on Investments in Individual Common Stocks." *Rev. Econ. and Stats.* 50 (August 1968): 312–31.

Stigler, George J., and Becker, Gary S. "De Gustibus Non Est Disputandum." *A.E.R.* 67 (March 1977): 76–90.

Book Reviews

Foreign Trade Regimes and Economic Development: Anatomy and Consequences of Exchange Control Regimes. By JAGDISH N. BHAGWATI.
Cambridge, Mass.: Ballinger Publishing Co. (for National Bureau of Economic Research), 1978. Pp. xix+232. $16.50.

This book is one of two volumes synthesizing the results of the NBER's project on Foreign Trade Regimes and Economic Development. The country studies covered Turkey, Ghana, Israel, Egypt, the Philippines, India, South Korea, Chile, and Colombia. Taken together with the country studies undertaken as part of the OECD's Industry and Trade project, whose results were summarized in the well-known volume by Little, Scitovsky, and Scott (1970), development economists now have a wealth of analytical and empirical material from which to judge the past trade policies of developing countries, as well as to make firmer policy pronouncements about future trade policy.

Bhagwati's synthesis volume covers two of the three issues with which the project was concerned, namely, "the characterization of the anatomy of exchange control regimes, including the interaction between quantitative and price measures" (p. 2), and the relationship between different types of exchange-control regimes and economic efficiency both in its static and dynamic aspects. In its third dimension, the project attempted to discover if there was "something systematic about the forces that governed the likelihood of success in moving from a restrictive exchange control regime to a liberalized regime" (p. 3). This is the subject matter of the companion volume by Anne Krueger (which is not reviewed here).

Within the compass that Bhagwati has set for himself, he has succeeded in providing a rigorous analysis and scrupulously fair summary of the evidence that he and others think is relevant to the anatomy and consequences of exchange-control regimes.

The starting point for Bhagwati's taxonomy of alternative exchange-control regimes is the limitations of the traditional Marshallian proposition on the equivalence of tariffs and quotas when its implicit assumption that the supply and demand schedules for foreign exchange are given breaks down. The actual ways in which imports (and/or exports) are regulated therefore become important in determining the relative domestic prices of traded goods and, hence, the welfare effects of the trade regime. As Bhagwati's subsequent taxonomy of the regulation mechanisms observed in the country studies shows, they are a bewildering variety, and if nothing else his taxonomy should help the practitioner attempting to analyze a particular LDC's trade

[*Journal of Political Economy*, 1981, vol. 89, no. 4]
© 1981 by The University of Chicago

regime to organize his or her thoughts and, as such, is a valuable addition to the literature. The rest (and major part) of the book summarizes the evidence from the project (and elsewhere) on the consequences of exchange-control regimes.

Aside from documenting the reasons for and evidence on the illegal transactions that different types of exchange-control regimes have entailed, the bulk of Bhagwati's book (pp. 82–218) is concerned with the static and dynamic efficiency effects of exchange-control regimes. The project used estimates of domestic resource costs (DRCs) as indicators of static efficiency. Bhagwati provides a lucid discussion (pp. 88–90) of the statistical and analytical pitfalls in the use of DRCs as opposed to "shadow prices" as measures of the misallocation of given resources. But despite these shortcomings, he rightly argues that "there is no easy way out of the conclusion that the pattern of investment was less than optimal" (p. 91) under the project Phase Two (i.e., QR [quantitative restrictions] based) regimes. Moreover, estimates of social rates of return based on "shadow prices" for at least India (see Lal 1980) show that these conclusions are unlikely to be overturned by the use of more appropriate indicators of allocative efficiency.

The evidence for the NBER country studies thus provides further support for the major conclusion of the earlier OECD study on the misallocation of resources resulting from import-substituting QR regimes. Thus Bhagwati writes,

> the weight of the evidence in the countries in the Project then is in support of the hypothesis that exchange control regimes by cheapening the cost of intermediates and capital goods influence and distort both the choice of techniques within activities and the relative composition of outputs and investments among activities. . . . while part of the evidence is inferential in deducing effects from the observed institutional features of the exchange control regimes of the Phase II variety and only part is direct, the overwhelming bulk of it is consistent with the view that the static efficiency effects of such regimes have been adverse in practice. [Pp. 117, 121]

The evidence presented on the dynamic effects of alternative trade regimes is less conclusive, though Bhagwati states that the major difference between the NBER and OECD projects was "in analyzing in far greater depth the growth effects as well as the conditions determining a successful transition from restrictive to liberalized regimes" (p. 6). The latter aspect is not discussed in this review, as it is the subject matter of Krueger's volume.

There is a long chapter reviewing the possible analytical links (and evidence thereon) between domestic and foreign saving and alternative trade regimes. Bhagwati rightly concludes that the evidence from the country studies and other econometric exercises does not support the view that restrictionist exchange-control regimes "will, or are likely to contribute to increased domestic savings and/or to augmented capital formation. If anything, much of our evidence—at least on the domestic savings issue—suggests an opposite relationship" (p. 174).

The project studies also provide documentation for the expected positive response of exports to a reduction in the bias imposed by an import-substituting QR regime; and moreover, as Bhagwati rightly emphasizes, "the project results indicate strongly that it is really a shift to successful liberalization and therefore *continuing* liberalization that is critical to improved export performance on a sustained basis" (p. 184).

The evidence cited and assessed on various other possible dynamic effects on entrepreneurship, innovation, and technical change and income distribution is very inconclusive, though it does support the negative conclusion that there is little basis for the view that there are beneficial dynamic effects in these dimensions which outweigh the static efficiency costs of import-substituting QR regimes.

Similarly, while "the cross-sectional contrasts among the Project countries indicate strongly that the countries that have shifted to [liberalized] Phases IV and V have better growth performance than those that have continued in Phase II–type restrictive regimes" (p. 191), the evidence for the reasons cited for why this may be the case again appears to be inconclusive.

Bhagwati, like many other development economists, asymmetrically labels a movement from the neutral free-trade position as import substituting, while a movement back toward this point is labeled export promoting. In an age given to slogan mongering, I am not sure that this redefinition of a policy of promoting free trade (or "optimal" trade) is necessarily helpful, as it could lead people (who have not read Bhagwati with care) to believe that a policy of export subsidization above the level which restores the equality between the effective exchange rates on imports and exports was desirable. Like the Little-Scitovsky-Scott volume, Bhagwati's message, though more muted, is that both in theory and practice a movement toward free trade is desirable in most developing countries, and it would seem best to say so clearly and emphatically.

I was also troubled by Bhagwati's hesitancy in arriving, as he puts it, "at strong prescriptive conclusions in regard to many of the detailed questions that are examined" (p. 5). In particular, on the dynamic effects of trade he states that, while most "*general* propositions [concerning the links between trade, savings, and growth] are not sustainable, *individual* country experiences may and often do indicate the dominance of only a subset of the possible linkages and effects, thus sustaining stronger propositions in regard to their own experience" (p. 5). But this raises the methodological question of the nature of the evidence required to substantiate the causal links in a chain of deductive economic argument linking various inductions. For, if one accepts the view that the dynamic case for free trade is that it is, in Irving Kravis's (1970) phrase, "a handmaiden of growth," it is not clear what other sort of general evidence would need to be adduced to support this view, apart from the historical and therefore counterfactual evidence presented for each country in the NBER sample. Surely on this basis we should be able to form a general prescriptive judgment in the dispute that arose during the 1950s between Nurkse and his followers on the one side, and Haberler and Viner on the other, on the desirability of free trade for postwar developing countries (see the quote in Bhagwati, p. 206).

The Nurkseian skepticism about the desirability of free trade for contemporary LDCs was based on three implicit or explicit assumptions. First was a pessimistic assumption about the external obstacles to LDC export growth, so that a shortage of foreign exchange was perceived to be chronic and an independent limiting factor in economic development. Second, economic agents in LDCs were assumed to be congenitally shortsighted in basing their investment and production decisions on current static rather than emerging dynamic comparative advantage. Finally, it was assumed that bureaucrats and planners could both predict the country's dynamic comparative advantage

better than private agents and ensure that production and investment patterns conformed to it through detailed state intervention. The OECD and NBER studies for the various countries have documented the inadmissibility of all three sets of assumptions. The obstacles to LDC export growth were largely internal, not external; economic agents in these countries have reacted to the distorted set of incentives created by protectionist trade-control regimes much in the way that standard economic theory would predict; planners have often shown a lack of foresight which would have swiftly bankrupted any private agent.

It is therefore surprising to find Bhagwati writing that "the statistical evidence and analysis of this possible cause and others of asymmetrical advantage of the EP strategy is not yet available in anything like the degree that would be reasonably compelling" (p. 215). Given the largely simple econometric evidence that Bhagwati cites in his discussion of the dynamic effects of trade regimes, I suspect that this is the kind of "compelling" evidence he is seeking. However, it is doubtful whether it will ever be possible to find evidence for the truly dynamic effects of trade regimes from which the usual statistical inferences could be made, and for a very compelling reason: Economics, as Hicks (1979, p. 38) has put it, "is on the edge of science and of history." The historical aspects are particularly important for what he terms "sequential causality," which is the relevant notion of causality when we seek to understand the dynamic effects of trade regimes. But in studying such dynamic historical processes, the use of techniques of statistical inference may not be very useful, because "when we cannot accept that the observations along the time series available to us are independent or cannot by some device be divided into groups that can be treated as independent, we get into much deeper water. For we have then, in strict logic, no more than one observation, all of the separate items having to be taken together. We are left to use our judgment, making sense of what has happened as best we can, in the manner of the historian. Applied economics does then come back to history after all" (Hicks 1979, p. 112).

The great strength of the NBER and OECD country studies is that they do provide analytical economic histories of particular LDC's trade regimes. These can and should be used to provide judgments on the validity or otherwise of one or the other "view of the world," but it is probably a vain hope to expect that it will ever be possible to derive "general laws" like those of physics for these dynamic historical processes by the use of statistical inference. May the inconclusiveness of Bhagwati's discussion of the dynamic effects of alternative trade regimes perhaps be due to his setting himself an impossible standard of evidence for substantiating the view that free trade regimes are the "handmaidens of growth"?

But perhaps this is being unfair to Bhagwati, who has written a lucid and scholarly book which, in his words, "the reader should approach . . . as a manual. Anyone who wishes to examine a specific country's foreign trade regime and its consequences can read the present volume and readily have access to possible hypotheses and techniques of analysis . . . that may be used for the reader's own analysis" (p. 5). In this he has succeeded and deserves commendation.

DEEPAK LAL

University College, London
World Bank

References

Hicks, John R. *Causality in Economics.* Oxford: Blackwell, 1979.

Kravis, Irving B. "Trade as a Handmaiden of Growth: Similarities between the Nineteenth and Twentieth Centuries." *Econ. J.* 80 (December 1970): 850–72.

Lal, Deepak. *Prices for Planning: Towards the Reform of Indian Planning.* London: Heinemann, 1980.

Little, Ian M. D.; Scitovsky, Tibor; and Scott, Maurice. *Industry and Trade in Some Developing Countries: A Comparative Study.* London: Oxford Univ. Press (for OECD Development Centre), 1970.

Population and Economic Change in Developing Countries. Edited by RICHARD A. EASTERLIN.
Chicago and London: University of Chicago Press, 1980. Pp. 581. $38.00.

This volume, number 30 of the Universities–National Bureau Committee for Economic Research conference reports, makes available papers and discussants' comments originally presented to the Conference on Population and Economic Change in Less Developed Countries held in Philadelphia in 1976. Nine distinguished authors have written major papers (a fact which must be stated when reviewing a conference volume, for, far too often, prominent researchers produce their second best for such occasions). Every paper is worth reading, and most are exciting. The volume will be necessary reading for people working in this and allied fields.

Yet the book may, in the years to come, be thought of as a monument in the final years of a profoundly unsatisfactory period—the Aristotelian era—in the study of the economic demography of developing countries. There are two major problems in the volume, and both are essentially problems of the discipline, faithfully reflected in the conference.

The first problem is that hardly any of the authors, and very few of the discussants, have ever spent any substantial period living and researching in a developing country. More limited experience is often just enough to convince the researcher, quite falsely, that he is in contact with a poorer version of our own society. It is as if anthropologists could still work without years of field experience, or as if geographers could write books from afar by merely interpreting local maps.

This is not just carping criticism. The central theme of both the conference and the book, addressed by five of the nine papers, is Third World household economics and individual behavior and decision making. These papers are saturated with hesitant inferences, drawn from piles of anonymous data collected in a society which the authors really do not know, about how these households must work. Very rarely does one of the writers say with any assurance that this or that kind of situation is usually found in the country being discussed. The loss is immense and can be illustrated by some related examples. The value of children's labor is compared with that of the adults in the family, or of outside paid labor, or that of slaves. No one points out that one reason for preferring children to adults is that they do various forms of menial work—"children's work"—that adults hate doing. No one points out that, throughout much of Africa and Asia, only family members are trusted, and the presence of hired labor in the house or even the surrounding fields

gives rise to disquiet. No one shows that children's labor is valuable only so long as they work hard with little resentment or protest, and that this is less a product of their existence than of a social system in which children expect to have their work directed by their elders and the older generation takes such direction for granted. It is a system which can be subtly transformed if cultural change (or the advent of schooling) alters the social relations between the generations. The market wage rate is frequently taken to measure the relative value (which must in a sense mean the intrinsic value and not be merely a synonym for market wage) of the labor of men, women, and children. Yet in a South Indian village in which my wife and I have lived and researched, although women are the preferred workers in certain postharvest operations because of the superior skills they are said to possess, their daily wage rate for that work is five-sevenths of that paid for the same work by men. Children's wages are lower still, even for tasks where they are demonstrably just as efficient as their elders. Wage levels do not measure work inputs, but differentials by sex and age are maintained in order to preserve a certain social fabric.

The second problem—related in the sense that it also arises from substituting pontification of a type for which we have condemned medieval schoolmen for long experience of the societies being analyzed—is the tendency to treat each society as autonomous. To a considerable extent this was the case in premodern Europe in that external cultures had no major impact, and hence Ronald Lee's perspective of preindustrial England escapes this censure. Yet it is not true in today's developing world. When we have asked people in both Africa and Asia about the major distinctions between their ways of life and those of their parents or more distant ancestors, they have far more often quoted aspects of social change than of economic development. Furthermore, they overwhelmingly explain such social change—and almost certainly correctly—in terms of imported Westernization (or social modernization) or because of learning other ways of life in school (seen quite clearly as an imported institution), rather than as the minimum alteration needed for accommodation to economic change. The social influence of the West, whether approved or disapproved, is felt to be immense. The decline of the patriarchy with all its economic advantages to older males is seen to have come about because of the new ideas. For most of the developing world the real fact of life is not a gradual development from the preexisting economic and social base but an accelerating revolution that has come about because of the extraordinary movement in this century toward a global economy (frequently noted in contexts other than this volume) and a global society (much less frequently noted). Yet almost every analysis in the book implies an almost autonomous economic and social change in every society, even tiny national states. That this is not really so can be inferred from some of the findings; for instance, Samuel Preston's proof that perhaps only one-third of the mortality decline this century in developing countries can be explained by rising income levels, while the other two-thirds (although it is not put this way) arises from the ever stronger interrelations between industrial and nonindustrial societies and indeed among all societies. This blind spot occurs not only because of inexperience of the Third World, but because the researcher becomes a prisoner of his own methods of analysis. The data for each country or society can easily be interrelated, but they cannot easily be made to divulge the extent to which change in their magnitudes is a result of external forces. In addition, the treatment of each society as autonomous allows apparently meaningful comparisons over distance and time.

There is a feeling, at least in demographic circles, that the era when approaches of this type have been possible is drawing to a close. The International Review Group of Social Science Research on Population and Development (IRG) Report, a forthcoming report of the International Union for the Scientific Study of Population (IUSSP), and other individuals and organizations are all urging population researchers toward microstudies, anthropological demography, and saturation in specific cultures. We are in the situation anthropology found itself in 60 years ago when Malinowski in Britain and Boas in the United States were trenchantly stating the necessity for knowing the real world. Economic theory has not crumbled through lack of such experience when it deals with the industrialized world only because most of its practitioners have known from a lifetime's experience the basic nature of that world. First World specialists in the population and economy of the Third World have not found themselves in such a fortunate position.

The first chapter takes the form of an excellent review and analysis by Peter Lindert of child costs from both contemporary and historical data. Yet today's Third World seems almost as far off as eighteenth-century England, and inferences about the Third World seem about as direct as today's inferences about eighteenth-century England.

The heart of the debate is provided by chapter 4, an attempt by Paul Schultz to apply the theories of the Chicago household economists to Taiwanese data, and chapter 5, a restatement by Richard Easterlin, Robert Pollak, and Michael Wachter of the University of Pennsylvania's tastes and preferences approach, this time written as a specific reaction to the household economists.

The application of the Chicago school's theories to developing countries has always been disconcerting because of its implication that the options for the use of time open to the middle-class American woman bear some relation to the situation of village women in the Third World. We have never been able to detect much difference in the allocation of time in the latter group between those who went to school and those who did not (ignoring the tiny fraction with higher education). The reason seems to be a cultural enforcement of the tasks which women must do and a sharing of child care between different members of the family and other relatives regardless of the mothers' level of education, plus a fairly rough-and-ready raising of children that often means little time spent exclusively on each child and little emotional drain. The differences that can be detected seem to run in the direction opposite that required by the Chicago school: A more educated mother can sometimes exert enough influence to increase the amount of time she spends with her children (and to insist that they are more exclusively her children) at the expense of more onerous work in the fields or in food processing. The application of the theories to Taiwan is particularly disconcerting for two reasons. The first is the now customary use of past education for current value of time in a society where education probably has far greater direct impact on age at marriage and use of contraception (as is shown by Ronald Freedman in his comments) and a profound impact on family relations and hence the ability to profit from family members' labor. It is high time that the Chicago school gave itself a basis for working in the Third World by encouraging studies at the household level of real time options and the making of choices. Such studies of 20 households would be as worthwhile as the comparison of thousands with macroscopic data employing probably wholly inappropriate proxies. The second, one readily conceded by Schultz, is the fact that comparisons could not be made with population units smaller than provinces (which, as Freedman pointed out, tend to be either city or noncity).

The Pennsylvania paper makes the point, which is universal knowledge in the Third World but is often ignored in the economic models of the First World so as to make them apparently universal, that before the onset of fertility transition there is very rarely any worry at all about the level of marital fertility (except when it is too low) and no greater attempts are made to restrict fertility at higher than at lower parity levels (a situation known somewhat unhappily as "natural fertility"). The age at marriage or the duration of sexual abstinence has nothing to do with intended ultimate family size. Indeed, decision making about family size does not exist and therefore cannot be studied until the onset of fertility transition—a forbidding situation for economic demographers because it rules out the universality of most of their models and the ability of these models to explain not only the conditions of stable high fertility but also the situation immediately preceding (and crucial in determining the timing of) the beginning of transition. The Pennsylvania trio makes the point that tastes and preferences must exist, but it seems to regard them as having some kind of life of their own rather than having underlying determinants (this is almost possible if you can convince yourself that the wobbles in the Western fertility decline that produced the so-called baby boom of the 1950s, and the falling fertility of the 1970s, are important phenomena in the transformation of stable high-fertility societies into low-fertility ones). They make many good points in producing a model so general that it borders on being a philosophic discussion rather than a tight description of the conditions of change (and, as Warren Sanderson charges, "the model is consistent with almost any kind of behavior").

Allen Kelley, employing a fraction of other people's data collected for other purposes in urban Kenya, comes to some fascinating conclusions: The number of children does not influence the rate of household financial saving; "the income effects of an increase in education may in some instances outweigh the price-substitution effects, resulting in an increase in family size with increasing education levels"; and (stating the obvious, although it is said to be a radical departure from the conventional wisdom) "contrary to the usual model where children are considered only as consumer durables, there is some evidence that children may also serve as producer durables in the low-income setting" (p. 436). The basic weakness of the paper is that the determination of savings by subtracting vastly greater expenditure figures from income figures, in a situation where neither is reliable, is wholly unconvincing.

Yoram Ben-Porath, in a paper which can be related only loosely to the major theme of the book, economic development and social change in a continuing society, examines the impact of the migration of Jews from Africa and Asia on their fertility and their replacement of dead children. His methods are ingenious, and he appears to show that "the income effects [in raising fertility] may precede in time the adjustment in life-style and aspirations that generates the fertility-depressing substitution effects" (p. 181). One problem is that more than life-style and aspirations changed: Economic realities both outside and inside the family did (the youth, who would work uncomplainingly in the family business in Yemen, wanted to seek an outside job in Israel and to control the spending of his earnings), as did the mortality of children. The other is the fact that the separation of demographic and other events between the pre- and postmigration periods is so difficult that one is not quite convinced that it has been wholly successfully achieved.

The volume does contain three fascinating macroscopic papers. One by Simon Kuznets argues (in a way that is necessary but should not be) that the mortality decline in developing countries over the last 60 years has not only

been a major contribution to real living standards but has disproportionately assisted the least privileged and thus has reduced social inequality. He argues that mortality reduction is an indispensable prerequisite for fertility decline and that we do not yet know the social impact of fertility decline in the Third World. He also shows that it is probable that the mortality declines in developing countries were as steep between the two world wars as since World War II. That is, they began when colonial administration was strong enough and sufficiently responsible to keep law and order and to apply the available health technology (as well as beginning the spread of schooling). It is a myth that the LDCs had to wait for antibiotics and WHO programs. Indeed, there is support for the widespread belief in the Third World and the historic West that it was strong regimes that benefited populations and that such benefit came in the form of prosperity and was evidenced by population growth arising from mortality decline.

There is evidence also (although it is not spelled out this way) in an important paper by Ronald Lee which extends his previous work. Mortality in England apparently declined consistently through the long years of the Tudor peace. The central theme of the analysis, however, is the apparent demonstration that population growth in England was for 600 years, from 1200 to 1800, controlled almost solely by changes in mortality levels (not themselves determined either economically or by variations in fertility) and that population growth so determined real wages that during the whole period a 1 percent movement in population gave rise to a 1.5 percent movement of wages in the opposite direction. The note of caution is sounded because there is less certainty about mortality trends for the whole population than for the upper classes and because the argument is most powerfully buttressed by swings in population between 1250 and 1450 which are far from certain.

Preston's paper, referred to earlier, also analyzes the mortality decline in terms of changing patterns in the cause of death and bravely attempts to work out a range of consequences of mortality decline.

Finally, the inclusion of Michael Todaro's paper on internal migration, clearly setting out his model and presenting a bibliographical guide to other studies, does something to justify the broad title of the book, though it appears somewhat anomalous in terms of the main thrust of the work.

The NBER volume presents, usually at one remove from the collection of data, much interesting material and a good deal of sophisticated analysis.

JOHN C. CALDWELL

Australian National University

The Regulated Industries and the Economy. By PAUL W. MACAVOY. New York: W. W. Norton & Co., 1979. Pp. 160. $11.95.

The regulatory economics field has grown rapidly since the publication of Alfred E. Kahn's (1970) seminal text 11 years ago. Powerful new institutions have been created to regulate health, safety, and environmental problems, several of the older industry-management agencies have relaxed their regulatory grip in response to active or threatened legislative deregulation, and the economics profession's understanding of the effects of regulation on the

economy has improved. Paul W. MacAvoy's latest book, *The Regulated Industries and the Economy,* evaluates both the success (or, more accurately, the failure) of the new social regulations and the increasing costs imposed by industry-management agencies in this recent inflationary period, concluding that the quality and the quantity of goods and services produced in the regulated industries have declined due to regulation.

The book aims at a general audience interested in problems of regulation but not necessarily well versed in economics. Its lack of coverage of many important topics in regulatory economics makes it unsuitable for a main text in the area. As a research monograph, the MacAvoy book is too loosely constructed to support satisfactorily many of its hypotheses. I found the data presented to suggest strongly some fascinating hypotheses, but often competing hypotheses were ignored, sufficiently refined data-analysis techniques were not used, and some broad, only weakly supported hypotheses were presented.

The book focuses on three main topics: the effects of price and entry regulation on the economy; the effects of health, safety, and environmental (social) regulation on the economy; and reform proposals. I found the evaluation of the impacts of price and entry regulation to be the most valuable section because its central hypothesis is important, as well as little recognized, and because that hypothesis is convincingly supported. The main finding states that due to their backward-looking, cost-based nature, the regulatory lag and money illusion that exist under price and entry regulation keep price increases behind and below cost increases, which causes a severe problem in high-inflation periods such as the 1970s (and, more important, the 1980s). This price squeeze pushes profit rates below average returns, thereby slowing investment and eventually leading to production shortages and reduced quality of service. MacAvoy illustrates his thesis with a careful analysis of the natural gas shortages that were caused by FPC price regulation, and he then persuasively argues that the electric power, railroad, and telephone industries may be headed along the same path. The seriousness of this result places reform of the regulatory process for these industries high on the political priority list.

Although reasonable evidence is presented to support most of the conclusions, many are overstated, without necessary qualifications, and without exploration of reasonable alternative hypotheses. As only one example, MacAvoy cites rate relief less than that requested by electric companies as evidence that regulation had some price-reducing effect, when the discrepancy between the requested and the granted rate increases can also be explained by gaming between both parties.

In contrast to his analysis of price and entry regulation, MacAvoy's analysis of the effects of social regulation is too simplistic. He studies three agencies, EPA, NHTSA, and OSHA, finds each of them to impose high costs and no significant benefits, and then concludes that social regulation is generally counterproductive.

The consistency of the conclusion may be satisfying to some, but it simply cannot be supported. The literature does suggest that OSHA and NHTSA have provided few benefits. However, the conclusion of MacAvoy's own study, that the EPA air programs have yielded no air-quality benefits, is difficult to accept, both because of questions about his methodology and because of its inconsistency with other studies. Even if OSHA, NHTSA, and the EPA air program did provide no benefits, it would not follow that other examples of social regulation provide benefits less than their costs or, worse

yet, provide no benefits at all. Although imposing severe costs on the economy in terms of forgone benefits from consumer products, the CPSC has also provided some benefits, if only from the added safety features it has induced manufacturers to adopt "voluntarily" to avoid CPSC action. And few knowledgeable about EPA water programs would argue that those programs have provided no significant improvements in water quality relative to what water quality would have been without the 1972 Water Pollution Control Act Amendments.

Regulatory reform comprises MacAvoy's third focus. He suggests reforms such as automatic rate increases (subject to later repayment if found to be excessive), the use of future test years, a requirement of economic impact statements for all significant regulations, and deregulation where competition, service quality, and the economy dictate this solution. His analysis is quite brief. I would like to see a book which thoroughly analyzes each of these and other reform proposals, compares them to existing procedures, and evaluates all their consequences to both industry and consumers.

Other than the problem of structuring a transition to competition in currently regulated "competitive" industries that will gain as many of the benefits of deregulation as possible, the most difficult regulatory problems in the next decade will consist of the reform of social regulation. For example, how should we implement Charles Schultze's (1977) proposal to replace command-and-control regulations with price incentives, such as marketable permits or effluent fees for environmental pollution? In rule making, how can we speed up the process while allowing open access to all interested parties and while guaranteeing that high-quality, unbiased scientific and economic analyses are used to support regulatory decisions? Given the rampant growth of regulation in the 1970s, can mechanisms like the existing Regulatory Council, the proposed regulatory budget, or the suggested legislative veto be expected to effectively coordinate many diverse and costly regulatory activities? These problems illustrate the need for careful thinking by economists about a variety of regulatory reforms that public policy must soon address.

WESLEY A. MAGAT

Duke University

References

Kahn, Alfred E. *The Economics of Regulation: Principles and Institutions*. 2 vols. New York: Wiley, 1970.

Schultze, Charles L. *The Public Use of Private Interest*. Washington: Brookings, 1977.

The Rational Peasant: The Political Economy of Rural Society in Vietnam. By SAMUEL L. POPKIN.
Berkeley: University of California Press, 1979. Pp. 306+xxi. $14.95 (cloth); $4.95 (paper).

The Origins of English Individualism: The Family, Property and Social Transition. By ALAN MACFARLANE.
New York: Cambridge University Press, 1979. Pp. 216+xv. $19.95 (cloth); $6.95 (paper).

Perhaps there is yet an economist who thinks peasants are irrational and collectivist. It is difficult to believe that such a benighted view of the countryside could persist, but this is the age of wonders. In truth, it remains a commonplace of development planning and scientific analysis. The two books under review here are signs of a counterattack.

Clear and important arguments can be described by their rhetorical dual as well as by their primal, by what they attack as well as by what they affirm. So here. As his title suggests, Samuel Popkin, a young political scientist at the University of California at San Diego, is attacking the World Bank view that peasants are idiots and the more sophisticated view, articulated recently by James C. Scott in *The Moral Economy of the Peasant* (1976), that peasants are rational but rationally communitarian. Popkin believes they are not communitarian in general and were not in particular over the past century in his country of study, Vietnam. He believes that the economic rationality of Vietnamese villagers in choosing crops was at one with their political rationality in choosing allies, both in the service of individual well-being.

The evidence for his beliefs is persuasive. The first third of the book is a long essay on "The Peasant," not necessarily Vietnamese, marshaling the latest work by medievalists and anthropologists (except Macfarlane's, published too late) against the views of Karl Polanyi, George Dalton, Eric Hobsbawm, Eric Wolf, James Scott, Joel Migdal, and Clifford Geertz. The "moral economists" "assume that peasants are antimarket, [that they] prefer common property to private, . . . [and] that peasant welfare depends on the closed corporate villages . . . . The fabric of the moral community is rent . . . by the intrusion of the 'cash nexus'" (pp. 5–6). The "political economy" view, as Popkin calls it, notes that the evidence for the moral economy view is weak. Peasants from Tepotzlan to Vilyatpur and from the 1280s to the 1970s, among other evidences of selfish and clearheaded maximizing, did not help each other to pay taxes (p. 41), demanded strict reciprocity in charity (p. 47), used gleaning rules to prevent misallocation of labor (not to help the poor [p. 52]), expected to grow rich out of village offices (p. 60), resisted innovation when it was rational to do so (p. 66), and engaged in exchanges of land or crops without social disaster (p. 68). The opening chapters constitute a brilliant survey of the context within which economists measure the supply response of African farmers and the transformation of traditional agriculture. A central planner who can read Popkin without some twinge of doubt that his bureaucracies, communes, and other substitutes for voluntary exchange are good for the poor (p. 29) is a hard man to persuade.

The rest of the book may do the trick. It is a narrative of Vietnamese political history down to 1946 emphasizing the impact of colonialism on the politics of villages. The narrative, like the analytical essay, is skillfully done, although of course it is difficult in any case for an American over 25 to yawn

at the names Hue, Tonkin, Cho Lon, and the like. The writing is modest, clear, and warm. So far as can be judged from his surefootedness in the literature on the European peasantry, Popkin's opinions in fields outside my competence are to be taken seriously. The index testifies to his heavy reliance on Blum and Hilton for Europe, and on Gourou, Gram, Ory, Osborne, Rambo, and Woodside for Vietnam. The book is generous to its enemies and almost embarrassing in its praise for allies. The footnotes are filled with references to studies that are "wonderful," "excellent," "must reading." The burden of bringing a chaotic history to bear on wider issues is sometimes too great, leading to excessive repetition and summarization. And the narrative portion can be faulted as rather too broadly focused to make the case about the details of village politics the argument requires. But the book is loaded with good observing and good thinking.

The major thought inspired by Popkin's history of Vietnam under the French is that the village was no passive victim of colonization and commercialization. On the contrary, village notables themselves called the new world into existence to redress the balance in the old. The new colonial taxes, for instance, were occasions for the notables to consolidate their power over the rest of the villagers (p. 149). Before the French came, "villages were not egalitarian, leveling, welfare oriented, or necessarily harmonious" (p. 132). In consequence, the many antifeudal movements of the twentieth century, whether religious or communist, did not look back to a Confucian golden age: the Vietnamese themselves knew well the rapacity of rural life. It is the analysts of the peasant wars of the twentieth century, not the warriors themselves, who idealize the communities the peasants came from.

Popkin criticizes too the other half of the idealized portrait of the death of community, namely, the vile and domineering colonialists who murdered it. The flaw, Popkin argues, is to attribute effectiveness in evil to the French administration. The French supported 10 times as many bureaucrats per head of local population as did the British in India yet failed repeatedly in their schemes to manipulate the villagers. The story of colonialism is not one of effective and evil government, which needs only to be replaced by effective and good government to reestablish the communitarian values of the golden age. It is a story of ineffective government lost in a sea of selfish peasants. The story should have a certain familiarity to Americans lately embroiled in Southeast Asia.

Popkin recognizes that the source of the mythology of the unselfish peasant is history, especially English history. So does Macfarlane, whose book is a more direct challenge to the mythology: "a great deal of modern India, Africa, or South America cannot be understood without some comprehension of what happened on a small island with only a few million inhabitants between the thirteenth and eighteenth centuries" (p. 7). The dual of Macfarlane's interpretation of English history is the one traditional since Marx, that England was like other economies long ago in having many "peasants." The peasant is here not merely a countryman but the denizen of a society with production in the home, no markets in products, no markets in factors of production, no geographical mobility, great emphasis on kinship, inheritance by the family instead of the individual, ownership by the family instead of the individual, an all-pervasive family life in which women marry young and are subordinate to their husbands, and sharp lines between this peasantry so described and the nobility and its state. In brief, the peasant is the man we all know and love from the works of Marx, Weber, Tawney, and their recent (collective) heirs—the same list that Popkin called the "moral economists."

The history that all educated people carry about in their heads is "a progression from small, isolated communities inhabited by 'peasants' . . . towards the market, monetized, 'open' social structure of the eighteenth century" (p. 54). The history, says Macfarlane, is wrong. England was in fact a nonpeasant society from earliest times for which we have records: there was no revolution of the sixteenth century. Like Popkin, he buttresses his own primary evidence (from two English villages that he himself is studying intensively) with a brilliant survey of work by others bearing on the point. No medievalist will find the facts reviewed novel, though he may be irritated by the single-mindedness with which they are deployed. Indeed, anyone who finds Macfarlane's (or Popkin's) theme entirely surprising has been relying for his view of English medieval society either on the works of Maitland or Vinogradoff, brilliant in their day (ca. 1900) but still in the grip of nineteenth-century German scholarship and its romantic view of the brotherhood of times past, or on textbooks by political historians who rely in turn on Maitland and Vinogradoff.

As you and I would turn to markets, and the political scientist Popkin turns to politics and power, it is natural that Macfarlane as an anthropologist turns to family relations to test the antiquity of English individualism. He argues that the English family was at an early stage nuclear, formed rather late in life, oriented toward the individual, respectful of women's rights, and unconfining. In particular, the individual, even women, inherited the family's land and could dispose of it at will. The central evidence is simply the records showing individuals disposing of land, in large amounts and with great freedom—as they did labor and produce. The notion that the early medieval (not to speak of late medieval) economy was a "natural economy," as the Germans used to put it, is flatly mistaken. *Sic transit* all manner of hypothetical histories, from Marx's notion of precapitalist economic formations, necessarily unsubstantiated in an age before the professionalization of history, down to such newer abominations, which may not be so easily pardoned, as Polanyi's *The Great Transformation* (1944), Hicks's *A Theory of Economic History* (1969), and Wallerstein's *The Modern World System* (1974).

Like Popkin, again, Macfarlane does not prove his case beyond all doubt. He is as uncomfortable with quantitative arguments as were the medieval men he studies and tries repeatedly to force the evidence to say yes or no when it can only say more or less. The most troublesome feature of the book is its insistence throughout that individualism was a peculiarity of England. What might be interpreted as a device of rhetoric to silence the cries of "it's-different-in-Dijon-and-Dacca" becomes in the end a disfigurement. Always it is "in England" or "at least in relation to England" that one after another myth of peasant life debunked, until finally (p. 199) we are told that by Adam Smith's time "England had probably long been different from almost every other major agrarian civilization." The evidence for this key assertion has the same weakness that Macfarlane detects throughout in others, as in the work of George Homans: "There appears to be a strong, self-confirming, and circular hypothesis. It is thought to be self-evident that the rural inhabitants of England [for present purposes read 'of any particular community'] were basically like 'peasants' elsewhere. . . . It is [then] justifiable to fill in the vast gaps . . . by drawing information from peasants elsewhere. The picture of [the society] which emerges then seems to show that there were really peasants" (p. 191).

For all his iconoclasm in the English case, then, Macfarlane retains a respectful attitude toward the notion that peasants elsewhere are a very odd

sort. The antidote is of course Popkin—Vietnam is most assuredly "elsewhere." From the union of these two books little remains of the myth of the peasant, historical or modern. Popkin's brief survey of the medieval evidence should be strengthened by Macfarlane; and Macfarlane's reluctance to extend his evidence outside England should be overcome by Popkin.

The books are not economics, though they talk much of economic matters. The question arises whether an economist should break off the scrutiny of his out-garbage from the comp center long enough to peruse a book on historical political science and a book on historical anthropology. Yes, he should. Historical mythologies live on in the premises of the economist who believes himself, whether proudly or humbly, to be a mere technician. To learn that "neither decline nor decay of peasant institutions is necessary for peasants to enter markets" (Popkin, p. 267) or that "'homo economicus' and market society [were] present in England for centuries" gives the economist new fields to conquer, and the conquered fields may in turn civilize his economics.

DONALD N. MCCLOSKEY

University of Iowa

References

Hicks, John R. *A Theory of Economic History*. Oxford: Clarendon, 1969.

Polanyi, Karl. *The Great Transformation*. New York: Farrar & Rinehart, 1944.

Scott, James C. *The Moral Economy of the Peasant. Rebellion and Subsistence in Southeast Asia*. New Haven, Conn.: Yale Univ. Press, 1976.

Wallerstein, Immanuel. *The Modern World System*. New York: Academic Press, 1974.

Rank-Order Tournaments as Optimum Labor Contracts

Edward P. Lazear and Sherwin Rosen

University of Chicago and National Bureau of Economic Research

This paper analyzes compensation schemes which pay according to an individual's ordinal rank in an organization rather than his output level. When workers are risk neutral, it is shown that wages based upon rank induce the same efficient allocation of resources as an incentive reward scheme based on individual output levels. Under some circumstances, risk-averse workers actually prefer to be paid on the basis of rank. In addition, if workers are heterogeneous in ability, low-quality workers attempt to contaminate high-quality firms, resulting in adverse selection. However, if ability is known in advance, a competitive handicapping structure exists which allows all workers to compete efficiently in the same organization.

I. Introduction

It is a familiar proposition that under competitive conditions workers are paid the value of their marginal products. In this paper we show that competitive lotteries are often efficient and sometimes superior to more familiar compensation schemes. For example, the large salaries of executives may provide incentives for all individuals in the firm who, with hard labor, may win one of the coveted top positions.

This paper addresses the relation between compensation and incentives in the presence of costly monitoring of workers' efforts and

We are indebted to Jerry Green, Merton Miller, James Mirrlees, George Stigler, Joseph Stiglitz, and Earl Thompson for helpful comments. The research was supported in part by the National Science Foundation. The research reported here is part of the NBER's research program in Labor Studies. Any opinions expressed are those of the authors and not those of the National Bureau of Economic Research.

[*Journal of Political Economy*, 1981, vol. 89, no. 5]
© 1981 by The University of Chicago 0022-3808/81/8905-0006$01.50

output. A wide variety of incentive payment schemes are used in practice. Simple piece rates, which have been extensively analyzed (see, e.g., Cheung 1969; Stiglitz 1975; Mirrlees 1976), gear payment to output. We consider a rank-order payment scheme which has not been analyzed but which seems to be prevalent in many labor contracts. This scheme pays prizes to the winners and losers of labor market contests. The main difference between prizes and other incentive schemes is that in a contest earnings depend on the rank order of contestants and not on "distance." That is, salaries are not contingent upon the output level of a particular game, because prizes are fixed in advance. Performance incentives are set by attempts to win the contest. We argue that in many circumstances it is optimal to set up executive compensation along these lines and that certain puzzling features of that market are easily explained in these terms.

Central to this discussion are the conditions under which mechanisms exist for monitoring productivity (Alchian and Demsetz 1972). If inexpensive and reliable monitors of effort are available, then the best compensation scheme is a periodic wage based on input. However, when monitoring is difficult, so that workers can alter their input with less than perfect detection, input-wage schemes invite shirking. The situation often can be improved if compensation is related to a more easily measured output level. In general, input-based pay is preferable because it changes the risk borne by workers in a favorable way. But when monitoring costs are so high that moral hazard is a serious problem, the gain in efficiency from using output-based pay may outweigh the risk-sharing losses. Paying workers on the basis of rank order alters costs of measurement as well as the nature of the risk borne by workers. It is for these reasons that it is sometimes a superior way to bring about an efficient incentive structure.

In the development below we start with the simplest case of risk neutrality to illustrate the basic issues. Then the more general case of risk aversion is treated in Section III. Section IV considers issues of sorting and self-selection when workers are heterogeneous.

II. Piece Rates and Tournaments with Risk Neutrality

To keep things simple and to avoid sequential and dynamic aspects of the problem, we confine attention to a single period in all that follows. Therefore, the reader should think of the incentive problem in terms of career development and lifetime productivity of workers. The worker's (lifetime) output is a random variable whose distribution is controlled by the worker himself. In particular, the worker is allowed to control the mean of the distribution by investing in costly skills

prior to entering the market. However, a given productivity realization also depends on a random factor which is beyond anyone's control. Employers may observe output but cannot ascertain the extent to which it is due to investment expenditure or to good fortune or to both, though workers know their input as well as output. Worker j produces lifetime output q_j according to

$$q_j = \mu_j + \epsilon_j, \tag{1}$$

where μ_j is the level of investment, a measure of skill or average output, chosen by the worker when young and prior to a realization of the random or luck component, ϵ_j. Average skill, μ_j, is produced at cost $C(\mu)$, with $C', C'' > 0$. The random variable ϵ_j is drawn out of a known distribution with zero mean and variance σ^2.[1] Here ϵ is lifetime luck such as life-persistent person-effects or an ability factor, which is revealed very slowly over the worker's lifetime. The crucial assumption is that productivity risk is nondiversifiable by the worker himself. That is another reason for choosing a long period for the analysis. For example, if the period were very short and the random factor was independently distributed across periods, the worker could diversify per period risk by repetition and a savings account to balance off good and bad years. Evidently a persistent person or ability effect cannot be so diversified when it is undiscoverable quickly, as appears true of managerial talent, for example. It is assumed, however, that ϵ is i.i.d. across individuals, so that owners of firms can diversify risk either by pooling workers together in one firm or by holding a portfolio.

To concentrate on incentive aspects of various contractual arrangements, we adopt the simplest technology for firms. Production requires only labor and is additively separable across workers. By virtue of the independence assumptions, managers act as expected value maximizers or as if they were risk neutral. Free entry and a competitive output market set the value of the product at V per unit. Again, these assumptions are adopted to illustrate basic issues in the simplest way. The analysis also applies when there are complementarities among workers in production, which is more realistic but more difficult to exposit.

Piece Rates

The piece rate is very simple to analyze when workers are risk neutral. It involves paying the worker the value of his product. Let r be the

[1] In this paper the worker has no choice over σ. This does not affect the risk-neutral solution but does have an effect if workers are risk averse, since they tend to favor overly cautious strategies. Also, virtually all the results of this paper hold true if the error structure is multiplicative rather than additive.

piece rate. Ignoring discounting, the worker's net income is $rq - C(\mu)$. Risk-neutral workers choose μ to maximize expected net return

$$E[rq - C(\mu)] = r\mu - C(\mu).$$

The necessary condition is $r = C'(\mu)$ or the familiar requirement that investment equates marginal cost and return. On the other hand, the expected profit of a firm is

$$E(Vq - rq) = (V - r)\mu,$$

so free entry and competition for workers imply $r = V$. Consequently

$$V = C'(\mu).$$

The marginal cost of investment equals its social return, yielding the standard result that piece rates are efficient.

Rank-Order Tournaments

We shall consider two-player tournaments in which the rules of the game specify a fixed prize W_1 to the winner and a fixed prize W_2 to the loser. All essential aspects of the problem readily generalize to any number of contestants. A worker's production follows (1), and the winner of the contest is determined by the largest drawing of q. The contest is rank order because the margin of winning does not affect earnings. Contestants precommit their investments early in life, knowing the prizes and the rules of the game, but do not communicate with each other or collude. Notice that even though there are two players in a given match the market is competitive and not oligopolistic, because investment is precommitted and a given player does not know who his opponent will be at the time all decisions are made. Each person plays against the "field."

We seek to determine the competitive prize structure (W_1,W_2). The method proceeds in two steps. First, the prizes W_1 and W_2 are fixed arbitrarily and workers' investment strategies are analyzed. Given these strategies, we then find the pair (W_1,W_2) that maximizes a worker's expected utility, subject to a zero-profit constraint by firms. It will be seen that a worker's incentives to invest increase with the spread between winning and losing prizes, $W_1 - W_2$. Each wants to improve the probability of winning because the return to winning varies with the spread. The firm would always like to increase the spread, *ceteris paribus*, to induce greater investment and higher productivity, because its output and revenue are increased. But as contestants invest more, their costs also rise. That is what limits the spread in equilibrium: Firms offering too large a spread induce excessive investment. A competing firm can attract all of these workers by decreasing the spread because investment costs fall by more than

expected product, raising expected net earnings. Increasing marginal cost of skill implies a unique equilibrium spread between the prizes that maximizes expected utility.

More precisely, consider the contestant's problem, assuming that both have the same costs of investment $C(\mu)$, so that their behavior is identical. A contestant's expected utility (wealth) is

$$(P)[W_1 - C(\mu)] + (1 - P)[W_2 - C(\mu)] = PW_1 + (1 - P)W_2 - C(\mu), \tag{2}$$

where P is the probability of winning. The probability that j wins is

$$\begin{aligned} P &= \text{prob}\,(q_j > q_k) = \text{prob}\,(\mu_j - \mu_k > \epsilon_k - \epsilon_j) \\ &= \text{prob}\,(\mu_j - \mu_k > \xi) = G(\mu_j - \mu_k), \end{aligned} \tag{3}$$

where $\xi \equiv \epsilon_k - \epsilon_j$, $\xi \sim g(\xi)$, $G(\cdot)$ is the cdf of ξ, $E(\xi) = 0$, and $E(\xi^2) = 2\sigma^2$ (because ϵ_j and ϵ_k are i.i.d.). Each player chooses μ_i to maximize (2). Assuming interior solutions, this implies

$$(W_1 - W_2)\frac{\partial P}{\partial \mu_i} - C'(\mu_i) = 0$$

and

$$(W_1 - W_2)\frac{\partial^2 P}{\partial \mu_i^2} - C''(\mu_i) < 0, i = j, k. \tag{4}$$

We adopt the Nash-Cournot assumptions that each player optimizes against the optimum investment of his opponent, since he plays against the market over which he has no influence. Therefore, j takes μ_k as given in determining his investment and conversely for k. It then follows from (3) that, for player j

$$\partial P/\partial \mu_j = \partial G(\mu_j - \mu_k)/\partial \mu_j = g(\mu_j - \mu_k),$$

which upon substitution into (4) yields j's reaction function

$$(W_1 - W_2)g(\mu_j - \mu_k) - C'(\mu_j) = 0. \tag{5}$$

Player k's reaction function is symmetrical with (5).

Symmetry implies that when the Nash solution exists, $\mu_j = \mu_k$ and $P = G(0) = \frac{1}{2}$, so the outcome is purely random in equilibrium. Ex ante, each player affects his probability of winning by investing.[2]

[2] However, it is not necessarily true that there is a solution because with arbitrary density functions the objective function may not be concave in the relevant range. It is possible to show that a pure strategy solution exists provided that σ^2 is sufficiently large: Contests are feasible only when chance is a significant factor. This result accords with intuition and is in the spirit of the old saying that a (sufficient) difference of opinion is necessary for a horse race. Stated otherwise, since $\partial P/\partial \mu_j = g(\mu_j - \mu_k)$ and $g(\cdot)$ is a pdf, $\partial^2 P/\partial \mu_j^2 = g'(\mu_j - \mu_k)$ may be positive, and fulfillment of second-order conditions in (4) implies sharp breaks in the reaction function. If σ^2 is small enough the breaks occur at very low levels of investment, and a Nash equilibrium in pure strategies will not exist Existence of an equilibrium is assumed in all that follows.

Substituting $\mu_j = \mu_k$ at the Nash equilibrium, equation (5) reduces to

$$C'(\mu_i) = (W_1 - W_2)g(0), \quad i = j, k, \tag{6}$$

verifying the point above that players' investments depend on the spread between winning and losing prizes. Levels of the prizes only influence the decision to enter the game, which requires nonnegativity of expected wealth.

The risk-neutral firm's realized gross receipts are $(q_j + q_k) \cdot V$, and its costs are the total prize money offered, $W_1 + W_2$. Competition for labor bids up the purse to the point where expected total receipts equal costs $W_1 + W_2 = (\mu_j + \mu_k) \cdot V$. But since $\mu_j = \mu_k = \mu$ in equilibrium, the zero-profit condition reduces to

$$V\mu = (W_1 + W_2)/2. \tag{7}$$

The expected value of product equals the expected prize in equilibrium. Substitute (7) into the worker's utility function (2). Noting that $P = ½$ in equilibrium, the worker's expected utility at the optimum investment strategy is

$$V\mu - C(\mu). \tag{8}$$

The equilibrium prize structure selects W_1 and W_2 to maximize (8), or

$$[V - C'(\mu)](\partial\mu/\partial W_i) = 0, \quad i = 1, 2. \tag{9}$$

The marginal cost of investment equals its marginal social return, $V = C'(\mu)$, in the tournament as well as the piece rate. Therefore, competitive tournaments, like piece rates, are efficient and both result in exactly the same allocation of resources.

Some further manipulation of the equilibrium conditions yields an interesting interpretation in terms of the theory of agency (see Ross 1973; Becker and Stigler 1974; Harris and Raviv 1978; and Lazear 1979):

$$\begin{aligned} W_1 &= V\mu + C'(\mu)/2g(0) = V\mu + V/2g(0) \\ W_2 &= V\mu - C'(\mu)/2g(0) = V\mu - V/2g(0). \end{aligned} \tag{10}$$

The second equality follows from $V = C'(\mu)$. Now think of the term $C'(\mu)/2g(0) = V/2g(0)$ in (10) as an entrance fee or bond that is posted by each player. The winning and losing prizes pay off the expected marginal value product plus or minus the entrance fee. That is, the players receive their expected product combined with a fair winner-take-all gamble over the total entrance fees or bonds. The appropriate social investment incentives are given by each contestant's attempt to win the gamble. This contrasts with the main agency result, where the bond is returned to each worker after a satisfactory performance has

been observed. There the incentive mechanism works through the employee's attempts to work hard enough to recoup his own bond. Here it works through the attempts to win the gamble.

Comparative statics for this problem all follow from (9) and (10) once a distribution is specified. For example, if ϵ is normal with variance σ^2, then $g(0) = \frac{1}{2}\sigma\sqrt{\pi}$. It follows from (10) that the optimal spread varies directly with V and σ^2. While several other interesting observations can be made of this sort, we note a somewhat different but important practical implication of this general scheme. Even though the optimal prize structure determines expected marginal product through its effect on worker choice of μ and the zero-profit condition (7) implies that expected prizes equal expected productivity, nevertheless actual realized earnings definitely do not equal productivity in either an ex ante or ex post sense. Consider ex ante first. Since $\mu_j = \mu_k = \mu$, expected products are equal. Since $W_1 > W_2$ is required to induce any investment, the payment that j receives never equals the payment that k receives. It is impossible that the prize is equal to ex ante product, because ex ante products are equal. Nor do wages equal ex post products. Actual product is Vq rather than $V\mu$. But q is a random variable, the value of which is not known until after the game is played, while W_1 and W_2 are fixed in advance. Only under the rarest coincidence would $W_1 = Vq_j$ and $W_2 = Vq_k$.

Consider the salary structure for executives. It appears as though the salary of, say, the vice-president of a particular corporation is substantially below that of the president of the same corporation. Yet presidents are often chosen from the ranks of vice-presidents. On the day that a given individual is promoted from vice-president to president, his salary may triple. It is difficult to argue that his skills have tripled in that 1-day period, presenting difficulties for standard theory where supply factors should keep wages in those two occupations approximately equal. It is not a puzzle, however, when interpreted in the context of a prize. The president of a corporation is viewed as the winner of a contest in which he receives the higher prize, W_1. His wage is settled on not necessarily because it reflects his current productivity as president, but rather because it induces that individual and all other individuals to perform appropriately when they are in more junior positions. This interpretation suggests that presidents of large corporations do not necessarily earn high wages because they are more productive as presidents but because this particular type of payment structure makes them more productive over their entire working lives. A contest provides the proper incentives for skill acquisition prior to coming into the position.[3]

[3] If ϵ is a fixed effect, there is additional information from knowing the identity of winners and losers. The expected productivity of a winner is $\mu + E(\epsilon_j | q_j > q_k)$, while that of a loser is $\mu + E(\epsilon_j | q_j < q_k)$. In a one-period contest there is no possibility of

Comparisons

Though tournaments and piece rates are substantially different institutions for creating incentives, we have demonstrated the surprising result that both achieve the Pareto optimal allocation of resources when workers are risk neutral. In fact other schemes also achieve this allocation. For example, instead of playing against an opponent, a worker might be compared with a fixed standard $\bar{q}$, with one payment awarded if output falls anywhere below $\bar{q}$ and another, higher, payment awarded if output falls anywhere above standard. Attempting to beat the standard has the same incentive effects as attempting to beat another player. Using the same methods as above, it is not difficult to show that there are spread-standard combinations that induce Pareto optimum investments. Since all these schemes involve the same investment policy, and since average payout by the firm equals average product for all of them, they all yield the same expected rewards and, therefore, the same expected utility to workers.[4]

In spite of the apparent equality of these schemes in terms of the preferences of risk-neutral workers, considerations of differential costs of information and measurement may serve to break these ties in practical situations. The essential point follows from the theory of measurement (Stevens 1968) that a cardinal scale is based on an underlying ordering of objects or an ordinal scale. In that sense, an ordinal scale is "weaker" and has fewer requirements than a cardinal scale. If it is less costly to observe rank than an individual's level of output, then tournaments dominate piece rates and standards. On the other hand, occupations for which output is easily observed save resources by using the piece rate or standard, or some combination, and avoid the necessity of making direct comparisons with others as the tournament requires. Salesmen, whose output level is easily observed, typically are paid by piece rates, whereas corporate executives, whose output is more difficult to observe, engage in contests.

In a modern, complex business organization, a person's productivity as chief executive officer is measured by his effect on the profitability of the whole enterprise. Yet the costs of measurement for

taking advantage of this information. However, in a sequential contest with no firm-specific capital, the information would be valuable and would constrain subsequent wage payments in successive rounds through competition from other firms. It is not difficult to show that this does not affect the general nature of the bond-gamble solution. Alternatively, if the investment has firm-specific elements or firms adopt policies that bind workers to it (as in Lazear 1979), these restrictions do not necessarily apply.

[4] The level of the standard is indeterminate, since for any $\bar{q}$ a corresponding spread can be chosen to achieve the optimal investment. This is also true of contests among more than two players. With N contestants, the prizes of $N - 2$ of them are indeterminate. When risk neutrality is dropped, the indeterminacy vanishes in both cases.

each conceivable candidate are prohibitively expensive. Instead, it might be said that those in the running are "tested" by assessments of performance at lower positions. Realizations from such tests are sample statistics in these assessments, in much the same way that grades are assigned in a college classroom and IQ scores are determined. The point is that such tests are inherently ordinal in nature, even though the profitability of the enterprise is metered by a well-defined, cardinal ratio scale. It is in situations such as this that the conditions seem ripe for tournaments to be the dominant incentive contract institution.

Notice in this connection that the basic prize and piece-rate structures survive a broad class of revenue functions other than summable ones. Even if the production function of the firm includes complicated interactions involving complementarity or substitution among individual outputs, there exists the possibility of paying workers either on the basis of individual performance or by rank order. The revenue function itself can even involve rank-order considerations, and both possibilities still exist. For example, spectators at a horse race generally are interested in the speed of the winning horse and the closeness of the contest. Then the firm's (track) revenue function depends on the first few order statistics; yet the horses could be paid on the basis of their speed rather than on the basis of win, place, and show positions. Both methods would induce them to run fast.[5]

There has been very little treatment of the problem of tournament prize structure and incentives in the literature. Little else but the well-known paper by Friedman (1953) based on Friedman-Savage preferences for lotteries exists in economics. In the statistics literature there is an early paper by Galton (1902) that is worthy of brief discussion. Galton inquired into the ratio of first- and second-place prize money in a race of n contestants, assuming the prizes were divided in the following ratio:

$$W_1/W_2 = (Q_1 - Q_3)/(Q_2 - Q_3).$$

Here Q_1 is the expected value of the first- (fastest) order statistic, etc. While a moment's reflection suggests this criterion to be roughly related to marginal productivity, Galton proposed it on strictly a priori grounds. He went on to show the remarkable result that the ratio above

[5] The reader is reminded that throughout this section and the next workers are identical a priori and differ only ex post through the realization of ϵ. In the real world, where there is population heterogeneity, market participants are sorted into different contests. There players (and horses, for that matter) who are known to be of higher quality ex ante may play in games with higher stakes. If it can be accomplished, the sorting is by anticipated marginal products. In that sense, pay differences among contestants of known quality resemble the effect of a "piece rate." These issues are more thoroughly discussed below.

is approximately 3 when the parent distribution of speed is normal. Hence, this criterion results in a highly skewed prize structure. From what we know today about the characteristic skew of extreme value distributions, a skewed reward structure based on order statistics is less surprising for virtually any parent distribution. In the more modern statistical literature, the method of paired comparisons has tournament-like features. Samples from different populations are compared pairwise, and the object is to choose the one with the largest mean. Comparing all samples to each other is like a round-robin tournament. An alternative design is a knockout tournament with single or double elimination. The latter requires fewer samples and is therefore cheaper, but does not generate as much information as the round robin (David 1963; Gibbons, Olkin, and Sobel 1977).

Galton's original work and the more modern developments it has given rise to are not helpful to us; they deal with samples from fixed populations, so the reward structure is irrelevant for resource allocation. The problem we have treated here is that of choosing the reward structure to provide the proper incentive and elicit the socially proper distributions.

III. Optimal Compensation with Risk Aversion

All compensation systems can be viewed as schemes which transform the distribution of productivity to a distribution of earnings. A piece rate is a linear transformation of output, so the distribution of income is the same apart from a change in location and scale. A tournament is a highly nonlinear transformation: It converts the continuous distribution of productivity into a discrete, binomial distribution of income. When workers are risk neutral, both schemes yield identical investments and expected utility because their first moments are the same. In this section, it is shown that with risk aversion one method or the other usually yields higher expected utility, because the interaction between insurance and action implies substantially different first and second moments of the income distribution in the two cases.[6]

We have been unable to completely characterize the conditions under which piece rates dominate rank-order tournaments and vice versa, but we show some examples here. Truncation offered by prizes implies more control of extreme values than piece rates but less control of the middle of the distribution. Different utility functions

[6] One might think that risks could be pooled among groups of workers through sharing agreements, but that is false because of moral hazard. A worker would never agree to share prizes since doing so would result in $\mu = 0$, and consequently $E(q_j + q_k) = 0$ and bankruptcy for the firm. As a result, firms offering tournaments or piece rates in the pure sense yield higher expected utility than the sharing arrangement.

weight one aspect more than the other so that tournaments can actually dominate piece rates.

Optimum Linear Piece Rate[7]

The piece-rate scheme analyzed pays workers a guarantee, I, plus an incentive, rq, where r is the piece rate per unit of output. The problem for the firm is to pick an r, I combination that maximizes workers' expected utility

$$\max_{I,r} [E(U) = \max \int U(y)\theta(y)dy], \tag{11}$$

where

$$\begin{aligned} y &\equiv I + rq - C(\mu) \\ &= I + r\mu + r\epsilon - C(\mu) \end{aligned} \tag{12}$$

and $\theta(y)$ is the pdf of y.

The worker's problem is to choose μ to maximize expected utility given I and r. If $\epsilon \sim f(\epsilon)$, the worker's problem is

$$\max_{\mu} E(U) = \int U[I + r\mu + r\epsilon - C(\mu)]f(\epsilon)d\epsilon.$$

The first-order condition is

$$\frac{\partial E(U)}{\partial \mu} = \int [U'(y)][r - C'(\mu)]f(\epsilon)d\epsilon = 0,$$

which conveniently factors so that

$$r = C'(\mu). \tag{13}$$

Condition (13) is identical to the risk-neutral case, because ϵ is independent of investment effort, μ.

Assuming risk-neutral employers, $V\mu$ is expected revenue from a worker and $I + r\mu$ is expected wage payments. Therefore, the zero-profit market constraint is

$$V\mu = I + r\mu. \tag{14}$$

Solving (14) for I and substituting into (12), the optimum contract maximizes

$$\int U\{V\mu(r) + r\epsilon - C[\mu(r)]\}f(\epsilon)d\epsilon$$

with respect to r, where $\mu = \mu(r)$ satisfies (13). After simplification the

[7] The following is similar to a problem analyzed by Stiglitz (1975). A linear piece-rate structure is a simplification. A more general structure would allow for nonlinear piece rates (see Mirrlees 1976).

marginal condition is

$$[V - C'(\mu)]\frac{d\mu}{dr}EU' + E\epsilon U' = 0. \tag{15}$$

Since risk aversion implies $E\epsilon U' < 0$, (15) shows that $V > C'(\mu)$ in the optimum contract for risk-averse workers. This underinvestment is the moral hazard resulting from insurance $I > 0$ and $r < V$ implied by (15).

Using familiar Taylor series approximations to the utility function and a normal density for ϵ, the optimum is approximated by

$$\mu \doteq C'^{-1}\left(\frac{V}{1 + sC''\sigma^2}\right) \tag{16}$$

and

$$\sigma_y^2 \doteq \frac{V^2\sigma^2}{(1 + sC''\sigma^2)^2}, \tag{17}$$

where $s \equiv -U''/U'$ evaluated at mean income is the measure of absolute risk aversion. Investment increases (see [16]) in V and decreases in s, C'', and σ^2, because all these changes imply similar changes in the marginal piece rate r which influences investment through condition (13). The same changes in V, s, and C'' have corresponding effects on the variance of income (see [17]), but an increase in σ^2 actually reduces variance, if σ^2 is large, because it reduces r and increases I.[8]

Optimum Prize Structure

The worker's expected utility in a two-player game is

$$E(U) = P\{U[W_1 - C(\mu^*)]\} + (1 - P)\{U[W_2 - C(\mu^*)]\}, \tag{18}$$

where * denotes the outcome of the contest rather than the piece-rate scheme. The optimum prize structure is the solution to

$$\max_{W_1, W_2}\left(E(U^*) = \max_{\mu^*}\{P \cdot U[W_1 - C(\mu^*)] + (1 - P) \cdot U[W_2 - C(\mu^*)]\}\right) \tag{19}$$

subject to the zero-profit constraint

$$V\mu^* = PW_1 + (1 - P)W_2. \tag{20}$$

The worker selects μ^* to satisfy $\partial E(U)/\partial\mu^* = 0$. Since cost functions are the same and ϵ_j and ϵ_k are i.i.d., the Nash solution implies $\mu_j = \mu_k$

[8] Furthermore, $r \doteq V/(1 + sC''\sigma^2)$ and $I \doteq sV^2\sigma^2/(1 + sC''\sigma^2)^2$, so that $r = V$ and $I = 0$ in the case of risk neutrality ($s = 0$). All these approximations use first-order expansions for terms in $U'(\cdot)$ and second-order expansions for terms in $U(\cdot)$. The same is true of the approximations below for the tournament.

and $P = \frac{1}{2}$ as before. Then the worker's investment behavior simplifies to

$$C'(\mu^*) = \frac{2[U(1) - U(2)]g(0)}{U'(1) + U'(2)}, \tag{21}$$

where $U(\tau) \equiv U[W_\tau - C(\mu^*)]$ and $U'(\tau) \equiv U'[W_\tau - C(\mu^*)]$ for $\tau = 1, 2$. Equation (21) implies

$$\mu^* = \mu^*(W_1, W_2), \tag{22}$$

and the optimum contract (W_1, W_2) maximizes

$$E(U^*) = \tfrac{1}{2}\, U[W_1 - C(\mu^*)] + \tfrac{1}{2}\, U[W_2 - C(\mu^*)] \tag{23}$$

subject to (20), with $P = \frac{1}{2}$, and (22). Increasing marginal cost of investment and risk aversion guarantees a unique maximum to (23) when a Nash solution exists. Again, assuming a normal density for ϵ, second-order approximations yield

$$\mu^* \doteq C'^{-1}\left(\frac{V}{1 + sC''\sigma^2\pi}\right) \tag{24}$$

and

$$\sigma^2_{y^*} \doteq \frac{\pi V^2 \sigma^2}{(1 + \pi s C''\sigma^2)^2}, \tag{25}$$

where

$$y^* \begin{cases} = W_1 - C(\mu^*) \text{ if } q_j > q_k \\ = W_2 - C(\mu^*) \text{ if } q_j < q_k \end{cases}$$

and $\epsilon_j \sim N(0,\sigma^2), \epsilon_k \sim N(0,\sigma^2)$, and cov $(\epsilon_j, \epsilon_k) = 0$.[9] The comparative statics of (24) and (25) are similar to the piece rate (16) and (17) and need not be repeated.

Comparisons

Equations (16) and (24) indicate that investment and expected income[10] are lower for the contest than for the piece rate at given values of s. Moreover, for values of σ^2 in excess of $1/sC''\sqrt{\pi}$, the variance of income in the tournament is smaller than for the piece rate. This would seem to suggest that contests provide a crude form of insurance when the variance of chance is large enough, but the problem is significantly more complicated than that because there is no separation between tastes and opportunities in this problem: The optimum

[9] Futhermore, $C'(\mu^*) \doteq g(0)(W_1 - W_2)$, so the spread is still crucial for investment incentives, as in the risk-neutral case.

[10] Since $y = V\mu - C(\mu)$, and since μ is below the wealth-maximizing level of μ when workers are risk averse, lower μ implies lower y because revenue falls by more than cost.

mean and variance themselves depend on utility-function parameters. Thus, for example, for the constant, absolute risk-aversion utility function $U = -e^{-sy}/s$, the insurance provided by the contest is insufficient to compensate for its smaller mean: It can be shown that the expected indirect utility of the optimal piece rate exceeds that of the optimal tournament for all values of σ^2, at least with normal distributions and quadratic investment-cost functions. However, when there is declining absolute risk aversion, we have examples where the contest dominates the piece rate.

Illustrative calculations are shown in table 1 using the utility function $U = \alpha y^{\alpha}$, which exhibits constant relative but declining absolute risk aversion, $s(y) = (1 - \alpha)/y$. Again quadratic costs and normal errors are assumed. However, this utility function is defined for positive incomes only, so an amount of nonlabor income y_0 is assigned to the worker to avoid a major approximation error of the normal, which admits negative incomes (i.e., the possibility of losses).

Table 1 shows that when $y_0 = 100$ so that $s = .005$, the contest is preferred until $\sigma^2 \geqslant 3$. However, if $y_0 = 25$ so that $s = .020$, the contest is only preferred for $\sigma^2 < .2$. The intuition is that piece rates concentrate the mass of the income distribution near the mean, while contests place 50 percent of the weight at one value significantly below the mean and the other value significantly above. Strongly risk-averse workers seem to dislike the binomial nature of this distribution when σ^2 is high because it concentrates too much of the mass at low levels of utility. However, when σ^2 is small, the contest which truncates the tails

TABLE 1

CONSTANT RELATIVE RISK AVERSION

σ^2	μ	μ^*	$E(U)$	$E(U^*)$
		$y_0 = 100$, $s(y_0) = .005$		
1	.9995	.9984	5.012155	5.012465
.5	9975	9922	5.012150	5.012445
1	.9950	.9846	5.012100	5.012295
3	.9852	.9552	5.011940	5.011925
6	9710	.9142	5.011800	5.011415
12	.9436	.8420	5.011420	5.010515
		$y_0 = 25$; $s(y_0) = .020$		
.1	.9980	.9938	2.524665	2.524725
.2	.9960	.9878	2.524616	2.524575
1	.9807	.9419	2.524237	2.523437
12	.8094	.5741	2.519930	2.514282

NOTE.—$U = \alpha y^{\alpha}$, $y = y_0 + I + rq - C(\mu)$ for piece rate; $y = y_0 + W_i - C(\mu)$ for contest ($i = 1, 2$); $\alpha = 5$, $V = 1$, $C(\mu) = \mu^2/2\ \sigma^2$

of the income distribution associated with a linear piece rate has higher value.

Income Distributions

While it is not possible to make a general argument based on an example, table 1 suggests that persons with more endowed income and smaller absolute risk aversion are more likely to prefer contests, and those with low levels of endowed wealth and larger absolute risk aversion are more likely to prefer piece rates. Consider a situation in which all persons have the same utility function, such as the one in table 1, and face the same costs and luck distribution, the only difference being the fact that some workers have larger endowed incomes than others. If this difference is large enough, it can be optimal to pay piece rates to those with small values of endowed income and to pay prizes to those with large values. Individuals will self-select the payment scheme in accordance with their wealth. The distribution of earnings among those selecting the piece-rate jobs is normal with mean $V\mu$ and variance $r^2\sigma^2$. It is binomial with mean $V\mu^*$ and variance $(\Delta W)^2/4$ for those who enter tournaments. Note that μ and μ^* depend upon $s(y)$, which is smaller for workers who select contests, and it can turn out as it does in table 1 that expected income is larger in the contest than in the piece rate; for example, if $\sigma^2 = 1$ then the rich prefer contests ($5.012295 > 5.012100$) and the poor prefer piece rates ($2.524237 > 2.523437$), but $\mu^* = .9846$ exceeds $\mu = .9807$. This situation is shown in figure 1.

The overall distribution is the sum of a binomial and a normal with lower mean, weighted by the number of individuals in each occupation (see fig. 1). It is positively skewed because $V\mu^* > V\mu$. Note also that the distribution of wage income will be less skewed than that of total income. The reason is that y_0 and mean-wage income are positively correlated because the likelihood of choosing a contest increases with y_0. These implications conform to the standard findings on the distribution of income in an economy.

This example is interesting because it is very closely related to some early results of Friedman (1953), who studied how alternative social arrangements can produce income distributions that cater to workers' risk preferences. He showed that the Friedman-Savage utility function leads to a two-class distribution. Persons in the risk-averse region are assigned to occupations in which income follows productivity, while persons in the risk-preferring region buy lottery tickets in very risky occupations in which few win very large prizes. The overall distribution is the sum of these two and exhibits characteristic skew. The Friedman-Savage utility function implies that a person's risk

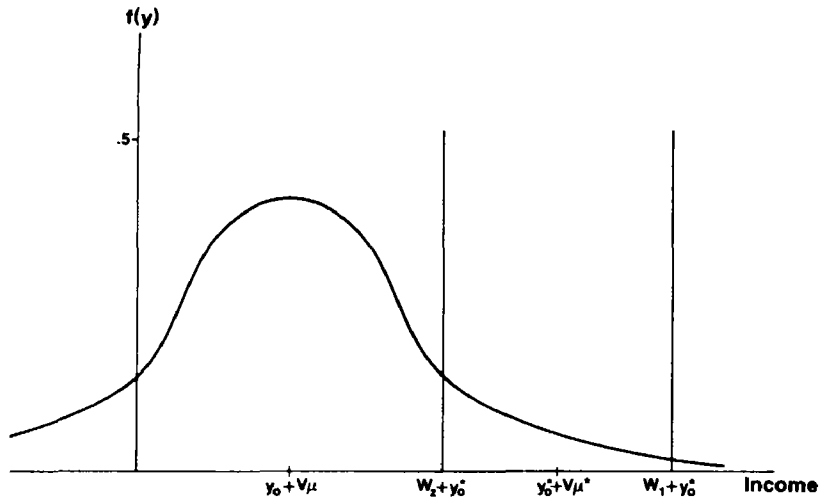


FIG. 1

preferences depend on the part of his wealth that is not at risk. Therefore, Friedman's assignment of people to jobs really follows endowed wealth (y_0), just as in our example. However, our framework offers two improvements. First, the problem of incentives is directly incorporated into the formulation of the optimum policy. Second, workers in this model are risk averse for all values of incomes, but even so gambles can be the optimal policy.

Error Structure

Relative costs of measurement are still important in choosing among incentive schemes, but the error structure plays additional roles when workers are risk averse. Suppose the output estimator for worker i in activity τ is $\hat{q}_{i\tau} = q_{i\tau} + \rho_\tau + \nu_{i\tau}$, where $\nu_{i\tau}$ is random error and ρ_τ is an error that is specific to activity τ but common to all workers within that activity. In the piece rate the common error ρ adds noise which risk-averse workers dislike, while the common noise drops out of a rank-order comparison because it affects both contestants similarly. That is, the relevant variance for the contest is $2\sigma_\nu^2$, while that for the piece rate is $\sigma_\rho^2 + \sigma_\nu^2$. It is evident that this can tip the balance in favor of tournaments if σ_ρ^2 is large enough and/or workers are sufficiently risk averse.

The common error ρ bears two interesting interpretations. One is activity-specific measurement error. For example, j and k may have the same supervisor whose biased assessments affect all workers similarly. This is similar to monitoring all workers by a mechanical count-

ing device that might run too fast or too slow in any given trial. The other interpretation of ρ is true random variation that affects the enterprise as a whole. For example, suppose all firms produce with the same technology, but that in a given period some firms do better or worse than others. Then risk-averse workers prefer not to have their incomes vary with conditions facing the firm as a whole, and wages based on a contest eliminate this kind of variation. Without its elimination there would be excessive losses due to moral hazard.

It must be pointed out that, in the absence of measurement error, using a contest against a fixed standard $\bar{q}$ discussed above has lower variance than playing against an opponent. As shown in Section II, the relevant variance in a contest is that of $\xi = \epsilon_k - \epsilon_j$, which has variance $2\sigma^2$ against an opponent and only σ^2 against a standard (since the standard is invariant, $\epsilon_k \equiv 0$). Consequently, we might expect risk-averse workers to prefer absolute standards.[11] Again, however, the crucial issue is the costs of measurement and the error structure. For the complex attributes required for managerial positions, it is difficult to observe output and therefore difficult to compare to an absolute standard. Insofar as samples and tests are necessary, it bears repeating that these are inherently ordinal in nature. But this leads us back to the problem of common error, where it is often impossible to know whether a person's output is satisfactory without comparisons to other persons. Further, when there are changing production circumstances in the firm as a whole, it is difficult to know whether the person failed to meet the standard because of insufficient investment or because the firm was generally experiencing bad times, a problem of measuring "value added." Risk-averse workers increase utility by competing against an opponent and eliminating this kind of firm effect.

IV. Heterogeneous Contestants

Workers are not sprinkled randomly among firms but rather seem to be sorted by ability levels. One explanation for this has to do with complementarities in production. But even in the absence of complementarities, sorting may be an integral part of optimal labor-contract arrangements. Informational considerations imply that

[11] Playing against a standard is like Mirrlees's (1976) notion of an "instruction." It is clear that using standards as well as piece rates must be superior to using one alone. That scheme would allow workers to be paid $\bar{I}$ if $q < \bar{q}$ and $I_0 + rq$ for $q \geq \bar{q}$. This is important because it truncates the possibilities when $Vq < 0$. Given the technology, it is possible that very large negative values of output can occur, and since it is impossible to always tax workers the full extent of this loss, some form of truncation is desirable. A contest is an alternative way to control the tails of this distribution.

compensation methods may affect the allocation of worker types to firms. Therefore, this section returns to the case of risk neutrality and analyzes tournament structures when investment costs differ among persons. Two types of persons are assumed, a's and b's, with marginal costs of the a's being smaller than those of the b's: $C_a'(\mu) < C_b'(\mu)$ for all μ. The distribution of disturbances $f(\epsilon)$ is assumed to be the same for both groups. Many of the following results continue to hold, with usually obvious modification of the arguments, if the a's and b's draw from different distributions. The following section addresses the question of self-selection when workers know their identities but firms do not. The next section discusses handicapping schemes when all cost-function differences can be observed by all parties.

Adverse Selection

Suppose that each person knows to which class he belongs but that this information is not available to anyone else. The principal result is that the a's and b's do not self-sort into their own "leagues." Instead, all workers prefer to work in firms with the best workers (the major leagues). Furthermore, there is no pure price-rationing mechanism that induces Pareto optimal self-selection. But mixed play is inefficient because it cannot sustain the proper investment strategies. Therefore, tournament structures naturally require credentials and other nonprice signals to differentiate people and assign them to the appropriate contest. Firms select their employees based on such information as past performances, and some are not permitted to compete.

The proof of adverse selection consists of two parts. First we show that players do not self-sort into a leagues and b leagues. Second, we show that the resulting mixed leagues are inefficient.

1. *Players do not self-sort.*—Assume leagues are separated and consider the expected revenue R_i generated by playing in league $i = a, b$ with an arbitrary investment level μ. Then

$$R_i(\mu) = W_2^i + (W_1^i - W_2^i)P^i, \quad i = a, b, \tag{26}$$

where (W_1^i, W_2^i) is the prize money, and P^i is the probability of winning in league i. Recall that P^i depends on the individual's level of investment and that of his rivals. Therefore, $P^a = G(\mu - \mu_a^*)$ and $P^b = G(\mu - \mu_b^*)$, where μ_a^* is the existing players' investments in the a league, where $V = C_a'(\mu_a^*)$, and similarly for μ_b^*. Recalling from (6) and (9) that $W_1^i - W_2^i = V/g(0)$ and from (10) that $W_2^i = V\mu_i^* - V/2g(0)$, equation (26) becomes

$$R_i(\mu) = V\mu_i^* - \frac{V}{g(0)}[\tfrac{1}{2} - G(\mu - \mu_i^*)]. \tag{27}$$

Note that $R_i(\mu) = V\mu_i^*$ when $\mu = \mu_i^*$ and that $dR_i/d\mu = R_i'(\mu) = Vg(\mu - \mu_i^*)/g(0) > 0$. Since $C_a'(\mu_a^*) = V$ and $C_b'(\mu_b^*) = V$, then $\mu_b^* < \mu_a^*$ so that $R_i(\mu_a^*) > R_i(\mu_b^*)$. Furthermore, $R_b'[\mu - (\mu_a^* - \mu_b^*)] = R_a'(\mu)$. Therefore, $R_b(\mu)$ is a pure displacement of $R_a(\mu)$. Since $R_i'(\mu) = V$ for $\mu = \mu_i^*$ and $R_i'(\mu) < V$ elsewhere, and since $R_i(\mu)$ is increasing, the revenue functions never cross. So $R_b(\mu)$ lies to the southwest of $R_a(\mu)$ (see fig. 2). Therefore, independent of cost curves, it is always better to play in the a league than the b league: Workers will not self-select.

2. *Mixed contests are inefficient.*—Suppose the proportions of a's and b's in the population are α and $(1 - \alpha)$, respectively. If pairings among a's and b's are random, then expected utility of a player of type i is

$$\overline{W}_2 + [\alpha P_a^i + (1 - \alpha)P_b^i](W_1 - W_2) - C_i(\mu_i),$$

where $(\overline{W}_1, \overline{W}_2)$ is the prize money in mixed play and P_j^i is the probability that a player of type i defeats a player of type j. The first-order condition for investment of type i in this game is

$$\left[\alpha \frac{\partial P_a^i}{\partial \mu_i} + (1 - \alpha) \frac{\partial P_b^i}{\partial \mu_i}\right] \cdot (\overline{W}_1 - \overline{W}_2) = C_i'(\overline{\mu}_i).$$

A development similar to Section II implies equilibrium reaction functions

$$[\alpha g(0) + (1 - \alpha)g(\overline{\mu}_a - \overline{\mu}_b)](W_1 - W_2) = C_a'(\overline{\mu}_a)$$

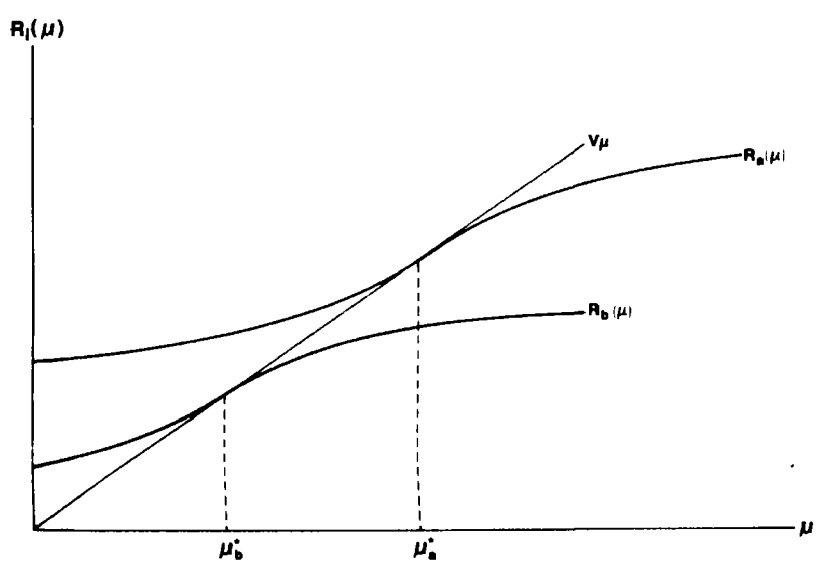


FIG. 2

for a's and

$$[\alpha g(\bar{\mu}_b - \bar{\mu}_a) + (1 - \alpha)g(0)](W_1 - W_2) = C_b'(\bar{\mu}_b)$$

for b's. If the solution is efficient, then $C_b'(\mu_b) = V = C_a'(\mu_a)$, which implies

$$\alpha g(0) + (1 - \alpha)g(\bar{\mu}_a - \bar{\mu}_b) = \alpha g(\bar{\mu}_b - \bar{\mu}_a) + (1 - \alpha)g(0).$$

Since g is symmetric and nonuniform, this condition can hold only if $\alpha = 1/2$. Therefore, except in that very special case, mixed contests yield inefficient investment: One type of player overinvests and the other underinvests depending upon whether or not $\alpha \gtrless 1/2$.

We conclude that a pure price system cannot sustain an efficient competitive equilibrium in the presence of population heterogeneity with asymmetric information. Markets can be separated, but only at a cost. Consider, for example, the case where a's want to prevent b's from contaminating their league. By making the spread, $W_1^a - W_2^a$, sufficiently large, $R_a(\mu)$ becomes steeper than $R_b(\mu)$ in figure 2 and crosses it so that the envelope covers $R_b(\mu)$ at low values of μ and $R_a(\mu)$ at high values. Then, for some high levels of μ, it is more profitable to play in the a league and, for low levels of μ, the b league is preferable. Individuals may self-sort, but the cost is that a's overinvest. The result is akin to that of Akerlof (1976) and to those of Spence (1973), Riley (1975), Rothschild and Stiglitz (1976), and Wilson (1977). As they show, a separating equilibrium need not exist, but, even if it does, that equilibrium may be inferior to a nonseparating equilibrium.

The obvious practical resolution of these difficulties is the use of nonprice rationing and certification to sort people into the appropriate leagues based on past performance. Similarly, firms use nonprice factors to allocate jobs among applicants. The rules for allocating those jobs may be important for at least two reasons that we can only briefly describe here.

First, sorting workers of different skill levels into appropriate positions within a hierarchy may be beneficial. In this paper, production is additive, so it does not matter who works with whom. To the extent that the production technology is somewhat more complicated, sorting may well be crucial. A series of pairwise, sequential contests may efficiently perform that function. Suppose that $q_{it} = \mu_i + \delta_i + \eta_{it}$, where δ_i is an unobserved ability component for player i and η is white noise. Suppose it is efficient for the individual with the highest δ to be the chief executive. There will be a tendency to have winners play winners because

$$E(\delta_j \mid q_{j1} > q_{k1}) > E(\delta_k \mid q_{j1} > q_{k1})$$

in the first round. A sequential elimination tournament may be a cost-efficient way to select the best person.

Second, workers may not know precisely their own abilities or cost functions. A worker who is ignorant about his cost function values information before selecting a level of investment expenditure. Therefore, firms may offer "tryouts" to provide information about optimal investment strategies. In fact, one can imagine the existence of firms which specialize in running contests among young workers—the minor leagues—which provide information to be used when and if the workers opt to increase the stakes and enter a bigger league.

These issues point up an important difference between piece rates and contests. In the pure heterogeneous case, where information is asymmetric and workers are risk neutral, a piece rate always yields an efficient solution, namely, $V = C_a'(\mu_a) = C_b'(\mu_b)$. However, once slotting of workers is important because of complementarities in production, or if it is desirable for workers to gain information about their type, it is no longer obvious that a series of sequential contests does not result in a superior allocation of resources.

Handicap Systems

This section moves to the opposite extreme of the previous discussion and assumes that the identities of each type of player are known to everyone. Competitive handicaps yield efficient mixed contests.

Consider again two types a and b now known to everyone. Prize structures in a-a and b-b tournaments satisfying (11) and (12) are efficient, but those conditions are not optimal in mixed a-b play. Denote the socially optimal levels of investment by μ_a^* and μ_b^*, their difference by $\Delta\mu$, and the prizes in a mixed league by $\bar{W}_1$ and $\bar{W}_2$. Let h be the handicap awarded to the inferior player b. Then the Nash solution in the a-b tournament satisfies

$$g(\mu_a - \mu_b - h)\Delta\bar{W} = C_a'(\mu_a) \tag{28}$$

and

$$g(\mu_a - \mu_b - h)\Delta\bar{W} = C_b'(\mu_b).$$

(The second condition in [28] follows from symmetry of $g[\xi]$.) Since the efficient investment criterion is $V = C_a'(\mu_a^*) = C_b'(\mu_b^*)$, independent of pairings, the optimum spread in a mixed match must be

$$\Delta\bar{W} = V/g(\Delta\mu - h). \tag{29}$$

From (28), condition (29) insures the proper investments by both contestants. The spread is larger in mixed than pure contests unless a

gives b the full handicap $h = \mu_a^* - \mu_b^*$. Otherwise, the appropriate spread is a decreasing function of h. Prizes $\bar{W}_1$ and $\bar{W}_2$ must also satisfy the zero-profit constraint $\bar{W}_1 + \bar{W}_2 = V \cdot (\mu_a^* + \mu_b^*)$ independent of h since the spread is always adjusted to induce investments μ_a^* and μ_b^*.

The gain to an a from playing a b with handicap h, rather than another a with no handicap, is the difference in expected prizes:

$$\begin{aligned}\gamma_a(h) &= \bar{P}\bar{W}_1 + (1 - \bar{P})\bar{W}_2 - C_a(\mu_a^*) - [(W_1^a + W_2^a)/2 - C_a(\mu_a^*)] \qquad (30)\\ &= \bar{P}\bar{W}_1 + (1 - \bar{P})\bar{W}_2 - (W_1^a + W_2^a)/2,\end{aligned}$$

where $\gamma_a(h)$ is the gain to a and $\bar{P} = G(\Delta\mu - h)$ is the probability that a wins the mixed match. The corresponding expression for b is

$$\gamma_b(h) = (1 - \bar{P})\bar{W}_1 + \bar{P}\bar{W}_2 - (W_1^b + W_2^b)/2. \qquad (31)$$

The zero-profit constraints in a-a, a-b, and b-b require that $\gamma_a(h) + \gamma_b(h) = 0$ for all admissible h. The gain of playing mixed matches to a is completely offset by the loss to b and vice versa.

If $C_a(\mu)$ is not greatly different from $C_b(\mu)$, then $\Delta\mu = \mu_a^* - \mu_b^*$ is small and $\bar{P} \doteq \frac{1}{2} + [g(\Delta\mu - h)](\Delta\mu - h)$. This approximation and the zero-profit constraint reduce (30) to

$$\gamma_a(h) \doteq V \cdot \left(\frac{\Delta\mu}{2} - h\right). \qquad (32)$$

The expression for $\gamma_b(h)$ is the same, except its sign is reversed, so the gain to a decreases in h, and the gain to b increases in h. Therefore, $h^* = \Delta\mu/2$ is the competitive handicap, since it implies $\gamma_a(h^*) = \gamma_b(h^*) = 0$. If the actual handicap is less than h^*, then γ_a is positive and a's prefer to play in mixed contests rather than with their own type, while b's prefer to play with b's only. The opposite is true if $h > h^*$.

A two-player game is said to be fair when the players are handicapped to equalize the medians. The competitive handicap does not result in a fair game, since $h^* = \Delta\mu/2 < \Delta\mu$. The a's are given a competitive edge in equilibrium, because they contribute more to total output in mixed matches than the b's do. This same result holds if ϵ_a has a different variance than ϵ_b, but it may be sensitive to the assumption of statistical independence and output additivity.

Alternatively, h can be constrained to be zero. In this case, different wage schedules would clear the market. Since $\gamma_a(0) = -\gamma_b(0) \equiv \beta$, paying $\bar{W}_1 - \beta$ and $\bar{W}_2 - \beta$ to a's, while paying $\bar{W}_1 + \beta$, $\bar{W}_2 + \beta$ to b's, leaves the spread and, therefore, the investments unaltered. It is easy to verify that a's and b's are still indifferent between mixed and pure contests, because expected returns are equal between segregated and integrated contests for each type of player. With no handicaps, the market-clearing prizes available to a's in the mixed contest are lower

than those faced by b's. Still, expected wages are higher for a's than b's in the mixed contest, because their probability of winning is larger. The b's are given a superior schedule in the mixed contest as an equalizing difference for having to compete against superior opponents. This yields the surprising conclusion that reverse discrimination, where the less able are given a head start or rewarded more lucratively if they happen to accomplish the unlikely and win the contest, can be consistent with efficient incentive mechanisms and might be observed in a competitive labor market.

V. Summary and Conclusions

This paper analyzes an alternative to compensation based on the level of individual output. Under certain conditions, a scheme which rewards rank yields an allocation of resources identical to that generated by the efficient piece rate. Compensating workers on the basis of their relative position in the firm can produce the same incentive structure for risk-neutral workers as does the optimal piece rate. It might be less costly, however, to observe relative position than to measure the level of each worker's output directly. This results in paying salaries which resemble prizes: wages which differ from realized marginal products.

When risk aversion is introduced, the prize salary scheme no longer duplicates the allocation of resources induced by the optimal piece rate. Depending on the utility function and on the amount of luck involved, one scheme is preferred to the other. An advantage of a contest is that it eliminates income variation which is caused by factors common to workers of a given firm.

Finally, we allow workers to be heterogeneous. This complication adds an important result: Competitive contests do not automatically sort workers in ways that yield an efficient allocation of resources when information is asymmetric. In particular, low-quality workers attempt to contaminate firms composed of high-quality workers, even if there are no complementarities in production. Contamination results in a general breakdown of the efficient solution if low-quality workers are not prevented from entering. However, when player types are known to all, there exists a competitive handicapping scheme which allows all types to work efficiently within the same firm.

References

Akerlof, George. "The Economics of Caste and of the Rat Race and Other Woeful Tales." *Q.J.E.* 90 (November 1976): 599–617.

Alchian, Armen A., and Demsetz, Harold. "Production, Information Costs, and Economic Organization." *A.E.R.* 62 (December 1972): 777–95.

Becker, Gary S., and Stigler, George J. "Law Enforcement, Malfeasance, and Compensation of Enforcers." *J. Legal Studies* 3, no. 1 (1974): 1–18.

Cheung, Steven N. S. *The Theory of Share Tenancy: With Special Application to Asian Agriculture and the First Phase of Taiwan Land Reform.* Chicago: Univ. Chicago Press, 1969.

David, Herbert A. *The Method of Paired Comparisons.* London: Charles Griffin, 1963.

Friedman, Milton. "Choice, Chance, and the Personal Distribution of Income." *J.P.E.* 61, no. 4 (August 1953): 277–90.

Galton, Francis. "The Most Suitable Proportion between the Values of First and Second Prizes." *Biometrika* 1 (1901–2): 385–90.

Gibbons, Jean D.; Olkin, Ingram; and Sobel, Milton. *Selecting and Ordering Populations: A New Statistical Methodology.* New York: Wiley, 1977.

Harris, Milton, and Raviv, Artur. "Some Results on Incentive Contracts with Applications to Education and Employment, Health Insurance, and Law Enforcement." *A.E.R.* 68 (March 1978): 20–30.

Lazear, Edward P. "Why Is There Mandatory Retirement?" *J.P.E.* 87, no. 6 (December 1979): 1261–84.

Mirrlees, James A. "The Optimal Structure of Incentives and Authority within an Organization." *Bell J. Econ.* 7 (Spring 1976): 105–31.

Riley, John G. "Competitive Signalling." *J. Econ. Theory* 10 (April 1975): 174–86.

Ross, Stephen A. "The Economic Theory of Agency: The Principal's Problem." *A.E.R. Papers and Proc.* 63 (May 1973): 134–39.

Rothschild, Michael, and Stiglitz, Joseph E. "Equilibrium in Competitive Insurance Markets: An Essay on the Economics of Imperfect Information." *Q.J.E.* 90 (November 1976): 630–49.

Spence, A. Michael. "Job Market Signaling." *Q.J.E.* 87 (August 1973): 355–74

Stevens, S. S. "Measurement, Statistics, and the Schemapiric View." *Science* 161 (August 30, 1968): 849–56.

Stiglitz, Joseph E. "Incentives, Risk and Information: Notes towards a Theory of Hierarchy." *Bell J. Econ. and Management Sci.* 6 (Autumn 1975): 552–79.

Wilson, Charles A. "A Model of Insurance Markets with Incomplete Information." *J. Econ. Theory* 16 (December 1977): 167–207.

An Exploration in the Theory of Exchange-Rate Regimes

Elhanan Helpman

Tel-Aviv University

Three exchange-rate regimes—a float, a one-sided peg, and a cooperative peg—are evaluated and compared in terms of efficiency and welfare levels. The framework of analysis embodies country-specific monies, with the money of each country being used to transact in its commodity markets and its currency-denominated bonds. Welfare levels depend only on consumption levels. In the presence of perfect foresight all equilibrium allocations are Pareto efficient. In a floating exchange-rate regime the perfect foresight equilibrium allocation coincides with an equilibrium of a costless barter economy. The same result holds in a one-sided peg if the pegging country's exchange-rate stabilizing authority breaks even over time. In a cooperative peg regime there is a different equilibrium allocation for each combination of exchange-rate levels and monetary policies. Problems of policy coordination and conflicts in desired monetary policies are discussed.

I. Introduction and Summary

The literature on international economics lacks systematic comparisons of welfare levels that are attained in different exchange-rate

This paper is based upon work supported by the National Science Foundation under grant no. SOC 78-06696 and the Foerder Institute for Economic Research. The main part of the study was performed when I was on leave at the University of Rochester. An earlier version of the paper was presented at seminars at the University of Rochester, the University of Western Ontario, the Institute for International Economic Studies at the University of Stockholm, and the University of Chicago. I wish to thank Robert Barro, Jacob Frenkel, Assaf Razin, and a referee for helpful comments.

[*Journal of Political Economy*, 1981, vol. 89, no. 5]
© 1981 by The University of Chicago. 0022-3808/81/8905-0004$01.50

systems. Nevertheless, this type of information is essential for a well-founded opinion about the relative desirability of alternative exchange-rate regimes, like fixed as opposed to floating. The need for research in this area can hardly be overemphasized, and it is more surprising that only very recently have there been attempts to provide such comparisons.

It is clear that in any analysis of exchange-rate systems the specification of the role of money will play a major role because, after all, an exchange rate is a relative price of two monies. It is, however, unsatisfactory to model money as an asset which serves only as a store of value, as is frequently done, because in this case money is just another asset without intrinsic features. An explicit modeling of these roles is required for a proper evaluation of the relative desirability of alternative exchange-rate regimes.

The present investigation is based on a specification of the role of money. Money can be used as a store of value, but it also has to be used in every market transaction. Purchases of commodities require payments in money, lending requires money to be transferred from the lender to the borrower, and debt repayments require money to be transferred from the debtor to the creditor. This is, of course, an application of the constraint transactions approach to the demand for money that was suggested by Clower (1967) (for a recent application see Lucas [1980*a*]). Most important, since I deal with more than one country—each country having its own money—it is assumed that for purchases of a particular country's goods one has to pay with that country's money. (This specification is also adopted by Stockman [1980], Helpman and Razin [1980], and Lucas [1980*b*].) Moreover, particular currency-denominated loans are given by providing the borrower with this currency. Debt repayments are done with the currency in which the debt is denominated. This specification of the transactions role of money introduces natural liquidity constraints into the economic system, and it clearly distinguishes money from other financial assets.

My framework of analysis, which incorporates money in the above-described way into an intertemporal model in which utility levels depend only on consumption levels, is presented in Section II. In the remainder of the paper the model is used to investigate three types of exchange-rate regimes: a floating exchange-rate regime and two types of fixed exchange-rate regimes. I discuss the implications of these regimes about global efficiency and about the distribution of welfare, as well as the determination of financial variables like prices, exchange rates, and interest rates. The emphasis is on financial aspects of exchange-rate regimes.

The fixed exchange-rate regimes that are considered are a one-sided peg and a cooperative peg. The one-sided peg is a regime in

which one country pegs its exchange rate vis-à-vis another currency, and the foreign exchange authority of the pegging country engages in foreign borrowing (lending) in order to finance its foreign exchange operations. The foreign exchange authority is constrained to repay its debts. In the cooperative regime there is a foreign exchange authority which is supported by all countries. This means that whenever it needs to sell a currency of a particular type, the country whose currency is required will provide the currency to the foreign exchange authority. There are important differences in the implications of these systems. At this stage I wish only to point out that, in the former, balance-of-payments adjustments do not change the quantity of money in circulation of the passive country but may change the quantity of money in circulation of the pegging country. This means that the value of money in circulation in the world economy may change as a result of balance-of-payments adjustments. In the latter system the aggregate value of money supply does not change as a result of balance-of-payments adjustments, which make it fall in line with recent modelings of the process of international adjustment under a fixed exchange rate (see, e.g., Mundell 1968, chap. 18).

Following the description of the framework of analysis, I discuss in Section III some general properties of equilibria. Most notably, it is shown that in the presence of perfect foresight the equilibrium allocation of consumption is Pareto efficient in every exchange-rate regime. The assumption of perfect foresight is extremely important for this result. It requires that governments specify their policies in advance and do not change them. For example, surprising monetary changes may destroy the efficiency of the resulting equilibrium. Although perfect foresight implies efficiency, the distribution of welfare depends on the exchange-rate system.

Section IV is devoted to a discussion of a floating exchange-rate regime. It is shown that in the presence of perfect foresight, and independently of monetary policies, the equilibrium allocation of consumption coincides with an equilibrium of a costless barter economy. Monetary policies which are specified in advance do not affect equilibrium consumption levels. It is clear from this result that in an equilibrium the liquidity constraints do not impose special constraints on the allocation of real resources. They do, however, help to determine nominal variables—prices, exchange rates, and interest rates—as is explained in that section.

The one-sided peg regime is discussed in Section V. It is shown that, provided the foreign exchange authority breaks even, in this regime too the resulting perfect foresight equilibrium allocation of consumption coincides with an equilibrium of a costless barter economy. Hence, a floating exchange-rate system and a one-sided peg have the same efficiency and distributional implications. There are,

however, also important differences. In particular, in the one-sided peg regime there is only one exchange-rate level which enables the foreign exchange authority to break even in the face of a given policy of monetary transfers (taxes). Hence, there is need for coordination of the exchange-rate level with the required accompanying monetary-fiscal policy mix. Coordination of policies is not required in a floating exchange-rate regime.

Section VI discusses the cooperative fixed exchange-rate regime. In this regime there is no need for policy coordination. But the perfect foresight equilibrium allocation of consumption and the distribution of welfare levels *depend* on the exchange-rate level and the monetary-fiscal policy mix that each country adopts. Only for very particular policy specifications will the cooperative peg equilibrium allocation of consumption coincide with the equilibrium allocation of consumption in a barter economy.

There are several interesting directions in which the present study can be extended in order to shed more light on the relative desirability of alternative exchange-rate regimes. Some of them are discussed in Section VII.

II. The Framework of Analysis

The world that I consider consists of two countries, a home (H) country and a foreign (F) country. Every country produces a single output, and the two outputs are perfect substitutes in consumption. In every period a country faces a given unalterable level of output. Time is discretely measured by time periods.

Neither the assumption of a single output nor the assumption of perfect substitutability in consumption is restrictive for the purpose of this paper; the main results do not depend on these assumptions. However, the assumption that output levels are unalterable does not enable me to discuss employment problems which are of major concern in discussions of exchange-rate systems (see, e.g., Friedman 1953; Flanders and Helpman 1978).

Every country has its own money. Purchases of the home country's goods are paid for with the home country's money; purchases of the foreign country's goods are paid for with the foreign country's money. Firms sell goods and transfer the profits to their stockholders. Home country firms are owned by home country residents while foreign country firms are owned by foreign country residents. Assuming that there are no costs of production, the value of a country's output in terms of its own money is transferred by firms to its residents.[1]

[1] Since I will not deal with uncertainty elements and I will assume the existence of

The precise specification of money flows is important. Money is needed for transacting in the market, and it imposes liquidity constraints on economic agents. One cannot buy goods unless one has the money to pay for them, and firms cannot distribute dividends unless they have the money to do so; all payments are done in the form of money. It is assumed, therefore, that firms distribute dividends *at the end* of each period. During a single period a firm accumulates money from sales of its goods, and it distributes this money to its owners at the end of the period.

An individual arrives at the beginning of period t with money balances that he left in his possession in period $t - 1$ (these may be home country as well as foreign country money balances) and with money balances that he has received from firms via dividend payments. At the beginning of period t he has to repay his short-term debts (debts may be positive or negative) that were acquired in period $t - 1$, including interest payments; these may be debts denominated in the home or foreign country's currency.

The individual can freely exchange home currency for foreign currency and vice versa. In a fixed exchange-rate regime this is done with the Exchange Rate Stabilizing Authority (ERSA for short), while in a floating exchange-rate regime this is done in the foreign exchange market. The individual can also borrow (lend) home or foreign country money at the beginning of period t. The completion of these transactions leaves him with certain quantities of home and foreign money as well as home and foreign currency–denominated debts (some of these may, of course, be zero). Now, part or all of the money balances can be spent on goods during period t—home currency on home goods and foreign currency on foreign goods—while the remaining part can be kept in order to increase next period's purchasing power. Thus, part of the total money holdings can be "money to spend" and part of it can be "money to hold" (see Hicks 1967).

In this paper I consider the case in which there are no restrictions on borrowing and lending behavior, except that debts have to be repaid. Foreign country decision variables are denoted by an asterisk, while home country decision variables are denoted without an asterisk. The world economy faces a horizon of length T—the same for both countries.[2]

bonds, there is no need to allow trade in ownership shares in firms, because such trade adds no new real trading opportunities.

[2] I assume a finite horizon. It is, however, easy to see from the following analysis that the results do not change when the horizon is infinite, provided equilibria exist. In order to analyze the infinite horizon case one has to replace constraints (1f) and (2f) with the requirement that the present value of debts is nonpositive for T going to infinity (see, e.g., Helpman and Razin 1981).

The home country's representative consumer solves the following problem:

Choose

$$\{c_{Ht}, c_{Ft}, M_{Ht}, M_{Ft}, B_{Ht}, B_{Ft}\}_{t=1}^{T} \tag{1}$$

to maximize

$$u_H(c_{H1} + c_{F1}, c_{H2} + c_{F2}, \ldots, c_{HT} + c_{FT}),$$

subject to

$$c_{Ht}, c_{Ft}, M_{Ht}, M_{Ft} \geq 0, \qquad t = 1, 2, \ldots, T, \tag{1a}$$

$$M_{H1} + e_1 M_{F1} \leq M_H + B_{H1} + e_1 B_{F1} + X_{H1}, \tag{1b}$$

$$p_{Ht} c_{Ht} \leq M_{Ht}, \qquad t = 1, 2, \ldots, T, \tag{1c}$$

$$p_{Ft} c_{Ft} \leq M_{Ft}, \qquad t = 1, 2, \ldots, T, \tag{1d}$$

$$\begin{aligned} M_{Ht} + e_t M_{Ft} \leq{}& (M_{Ht-1} - p_{Ht-1} c_{Ht-1}) + e_t (M_{Ft-1} - p_{Ft-1} c_{Ft-1}) \\ &+ p_{Ht-1} y_{Ht-1} - (1 + i_{Ht-1}) B_{Ht-1} - (1 + i_{Ft-1}) \\ &\cdot e_t B_{Ft-1} + B_{Ht} + e_t B_{Ft} + X_{Ht}, t = 2, 3, \ldots, T, \end{aligned} \tag{1e}$$

$$B_{Ht}, B_{Ft} \leq 0, \tag{1f}$$

where

u_H = home country utility function that is strictly increasing in all arguments,
c_{Ht} = consumption of home output in period t,
c_{Ft} = consumption of foreign output in period t,
M_{Ht} = beginning-of-period holdings of home money in period t,
M_{Ft} = beginning-of-period holdings of foreign money in period t,
B_{Ht} = home currency borrowing (lending when negative) at the beginning of period t,
B_{Ft} = foreign currency borrowing (lending when negative) at the beginning of period t,
M_H = the initial stock of home country money,
X_{Ht} = government transfers (taxes if negative) to home country residents in terms of home money at the beginning of period t,
e_t = the exchange rate in period t, defined as the price of foreign currency in terms of the home country's currency,
p_{Ht} = the price of home output in terms of home currency in period t,
p_{Ft} = the price of foreign output in terms of foreign currency in period t,
y_{Ht} = home output in period t,

i_{Ht} = the nominal one-period interest rate on period t home currency–denominated debt, and

i_{Ft} = the nominal one-period interest rate on period t foreign currency–denominated debt.

At the beginning of the first period, home country residents have M_H units of home currency, and they receive a transfer of X_{H1} units of home currency from their government. They borrow B_{H1} units of home money and B_{F1} units of foreign money. Then, they reallocate total money holdings between home and foreign currency via the foreign exchange market in a floating exchange-rate regime or the ERSA in a fixed exchange-rate regime. This way they end up holding M_{H1} units of home money and M_{F1} units of foreign money. No further financial transactions are needed during the first period. Consumers spend part or all of their money holdings on goods during the period. Thus, total beginning-of-period home money holdings are composed of $p_{H1}c_{H1}$ units of money to spend and $M_{H1} - p_{H1}c_{H1}$ units of money to hold. This is true for foreign money holdings as well.

At the beginning of the second period, home consumers have $M_{H1} - p_{H1}c_{H1}$ units of home currency and $M_{F1} - p_{F1}c_{F1}$ units of foreign currency. They receive dividend payments which consist of $p_{H1}y_{H1}$ units of home currency and government transfers which consist of X_{H2} units of home currency. They repay their first-period debts, including interest payments—$(1 + i_{H1})B_{H1}$ in home currency and $(1 + i_{F1})B_{F1}$ in foreign currency—and acquire new debts—B_{H2} in home currency and B_{F2} in foreign currency. They can make instantaneous transactions in the foreign exchange market or with the ERSA in order to coordinate the momentary money flows. Finally, they reallocate total money holdings between home and foreign money—M_{H2} and M_{F2}. And the process continues.

We may now turn to the decision problem of the foreign country's representative consumer. His decision variables are denoted by an asterisk, the initial stock of foreign money is denoted by M_F, foreign output in period t is denoted by y_{Ft}, and government transfers (taxes if negative) to foreign residents in terms of foreign money at the beginning of period t are denoted by X_{Ft}. All other variables are as defined above (except for the asterisk).

Choose

$$\{c^*_{Ht}, c^*_{Ft}, M^*_{Ht}, M^*_{Ft}, B^*_{Ht}, B^*_{Ft}\}_{t=1}^{T} \tag{2}$$

to maximize

$$u_F(c^*_{H1} + c^*_{F1}, c^*_{H2} + c^*_{F2}, \ldots, c^*_{HT} + c^*_{FT}),$$

subject to

$$c^*_{Ht}, c^*_{Ft}, M^*_{Ht}, M^*_{Ft} \geq 0, \qquad t = 1, 2, \ldots, T, \tag{2a}$$

$$\frac{M^*_{H1}}{e_1} + M^*_{F1} \leq M_F + \frac{B^*_{H1}}{e_1} + B^*_{F1} + X_{F1}, \tag{2b}$$

$$p_{Ht}c^*_{Ht} \leq M^*_{Ht}, \qquad t = 1, 2, \ldots, T, \tag{2c}$$

$$p_{Ft}c^*_{Ft} \leq M^*_{Ft}, \qquad t = 1, 2, \ldots, T, \tag{2d}$$

$$\begin{aligned} \frac{M^*_{Ht}}{e_t} + M^*_{Ft} \leq{}& \frac{(M^*_{Ht-1} - p_{Ht-1}c^*_{Ht-1})}{e_t} + (M^*_{Ft-1} - p_{Ft-1}c^*_{Ft-1}) \\ &+ p_{Ft-1}y_{Ft-1} - (1 + \iota_{Ht-1})\frac{B^*_{Ht-1}}{e_t} - (1 + \iota_{Ft-1})B^*_{Ft-1} \\ &+ \frac{B^*_{Ht}}{e_t} + B^*_{Ft} + X_{Ft}, t = 2, 3, \ldots, T, \end{aligned} \tag{2e}$$

$$B^*_{HT}, B^*_{Ft} \leq 0. \tag{2f}$$

This is the foreign counterpart of the home country's decision problem (1).

III. A Fundamental Property of Equilibria

The purpose of this section is to show that all perfect foresight equilibria, independently of the exchange-rate regime, are Pareto optimal. This is an extremely strong result because it implies that in the present framework the exchange-rate regime is irrelevant from the efficiency viewpoint; that is, there is no exchange-rate regime that is more efficient than other exchange-rate regimes. This does not mean, of course, that exchange-rate regimes are irrelevant from the viewpoint of other economic considerations. For example, it may happen that different exchange-rate regimes have different distributional implications in the sense that in some of them a particular country is better off than in others. I will address this issue in the following sections when discussing particular exchange-rate regimes. At this stage I should like only to point out that the equivalence result depends strongly on the ability of governments to pursue nondistorting absorption policies whenever they are needed in order to accompany an exchange-rate regime. I will have more to say about this in Section V.

First, independently of the exchange-rate system, in an equilibrium the interest rate and purchasing power parity conditions have to be satisfied. In order to have a solution to (1) and (2) with finite values of $B_{Ht}, B_{Ft}, B^*_{Ht}, B^*_{Ft}, t = 1, 2, \ldots, T - 1$ (last-period debt is always zero

due to [1f] and [2f]), there has to be interest rate parity in every period:

$$(1 + i_{Ht}) = (1 + i_{Ft})\frac{e_{t+1}}{e_t}, \quad t = 1, 2, \ldots, T - 1. \tag{3}$$

For if (3) does not hold, an investor can arbitrage between foreign and domestic debt so as to make infinite profits. One can use problems (1) and (2) to show that when $(1 + i_{Ht}) < (1 + i_{Ft})(e_{t+1}/e_t)$ the solutions to these problems yield B_{Ht}, $B^*_{Ht} \to +\infty$ and B_{Ft}, $B^*_{Ft} \to -\infty$, and when $(1 + i_{Ht}) > (1 + i_{Ft})(e_{t+1}/e_t)$ the solutions to these problems yield B_{Ht}, $B^*_{Ht} \to -\infty$ and B_{Ft}, $B^*_{Ft} \to +\infty$. These asset demands are, of course, inconsistent with an equilibrium.

Second, in order to have equilibrium in commodity markets there has to be purchasing power parity in every period:

$$p_{Ht} = e_t p_{Ft}, \qquad t = 1, 2, \ldots, T. \tag{4}$$

For if $p_{Ht} > e_t p_{Ft}$, the solutions to (1) and (2)—taking into account (3)—yield zero demand for the home country's output in period t, which is inconsistent with the clearing of commodity markets. Similarly, if $p_{Ht} < e_t p_{Ft}$, demand for foreign output drops to zero.

In order to prove the main result of this section, it is useful to rewrite the consumer problems so as to replace the sequence of budget constraints by a single wealth constraint. This can be done in the present context despite the existence of liquidity constraints, because (*a*) there exist bonds, (*b*) financial assets (including monies) are freely traded at the beginning of each period, and (*c*) the interest parity and purchasing power parity conditions are satisfied.

First, observe that in a solution to problem (1) constraints (1b) and (1e) will never be satisfied with strict inequality. For if there is a strict inequality in one of these constraints, say, in period t, one can increase domestic or foreign money holdings at no extra cost and use it as money to spend, thus increasing consumption in period t without altering all other choice variables. This will increase the attained utility level. We can rewrite, therefore, constraints (1b) and (1e) in equality form without altering the nature of the problem. Second, denoting by m_{Ht} and m_{Ft} the components of home and foreign monies to hold in period t,

$$m_{Ht} = M_{Ht} - p_{Ht}c_{Ht}, \qquad t = 1, 2, \ldots, T, \tag{5a}$$

$$m_{Ft} = M_{Ft} - p_{Ft}c_{Ft}, \qquad t = 1, 2, \ldots, T, \tag{5b}$$

we can replace (1c) and (1d) with nonnegativity constraints on m_{Ht} and m_{Ft}.

Combining (1b) and (1e)—written in equality form—with (5), and

applying the parity conditions (3) and (4), we get:

$$p_{H1}c_1 + m_{H1} + e_1 m_{F1} = M_H + B_{H1} + e_1 B_{F1} + X_{H1}, \tag{6}$$

$$\begin{aligned}(1 + i_{Ht-1})^{-1}p_{Ht}c_t + (1 + i_{Ht-1})^{-1}(m_{Ht} + e_t m_{Ft}) - (1 + i_{Ht-1})^{-1}(m_{Ht-1} \\ + e_t m_{Ft-1}) = (1 + i_{Ht-1})^{-1}(p_{Ht-1}y_{Ht-1} + X_{Ht}) - B_{Ht-1} \\ - e_{t-1}B_{Ht-1} + (1 + i_{Ht-1})^{-1}(B_{Ht} + e_t B_{Ft}), \qquad t = 2, 3, \ldots, T,\end{aligned} \tag{7}$$

where $c_t = c_{Ht} + c_{Ft}$ = aggregate consumption in period t.

Now define the discount factors:

$$d_{Ht} = \prod_{\tau=2}^{t}(1 + i_{H\tau-1})^{-1}, \qquad t = 2, 3, \ldots, T, \text{ and } d_{H1} = 1. \tag{8}$$

Then, by adding (7) for $t = 2$ to (6), adding the result to (7) for $t = 3$ multiplied by d_{H2}, then adding the result to (7) for $t = 4$ multiplied by d_{H3}, and so on, we get (using the interest parity condition [3]) the single wealth constraint:

$$\begin{aligned}\sum_{t=1}^{T} d_{Ht}p_{Ht}c_t + \sum_{t=1}^{T-1} d_{Ht+1}(i_{Ht}m_{Ht} + e_{t+1}i_{Ft}m_{Ft}) \\ + d_{HT}(m_{HT} + e_T m_{FT} - B_{HT} - e_T B_{FT}) \\ = M_H + \sum_{t=1}^{T-1} d_{Ht+1}p_{Ht}y_{Ht} + \sum_{t=1}^{T} d_{Ht}X_{Ht}.\end{aligned} \tag{9}$$

On the left-hand side of (9) there are three terms. The first represents the present value of consumption. The second represents the present value of interest costs associated with holding money as a store of value, that is, associated with money to hold. The last term represents the present value of last-period money to hold minus debt. On the right-hand side of (9) there are three terms too—initial money holdings, the present value of dividends, and the present value of government transfers. The sum of the right-hand-side terms represents initial wealth. Dividends are distributed at the end of each period, which explains why the firms' proceeds in period t are discounted by d_{Ht+1} and not by d_{Ht}.

Now problem (1) can be reformulated as follows. Choose $c_t, m_{Ht}, m_{Ft} \geq 0, t = 1, 2, \ldots, T$, and $B_{HT}, B_{FT} \leq 0$, so as to maximize $u_H(c_1, c_2, \ldots, c_T)$ subject to (9). Having found a solution to this problem, one can find a solution to problem (1) by choosing an arbitrary division of consumption in each period between home and foreign goods, then using (5) to solve beginning-of-period money holdings, and finally using (6) and (7) to solve home and foreign borrowing.

Applying to problem (2) the above-outlined procedure, one can derive the foreign counterpart of (9). Due to the interest and purchasing power parity conditions, the foreign wealth constraint can be written using foreign currency prices and discount factors or home currency prices and discount factors. The foreign country's decision problem can then also be reformulated using a single wealth constraint.

The new formulations of problems (1) and (2) reveal an important characteristic of equilibrium allocations. If the interest rate on currency j–denominated debt is positive in period t ($j = H$ or F), neither the home nor the foreign country will use currency j as a store of value in period t (i.e., m_{jt} and m^*_{jt} will be chosen to equal zero). If, however, the interest rate on currency j–denominated debt is zero in period t, both countries will be indifferent toward using currency j as a store of value in period t. This means that, in an equilibrium, implicit interest costs associated with money to hold are zero in every period.[3] This can be seen by observing that in the reformulated problems the "price" of home money to hold is $d_{Ht+1}i_{Ht}$ and the "price" of foreign money to hold is $d_{Ht+1}e_{t+1}i_{Ft}$, while the marginal utility of money to hold is zero. The new formulation reveals also that in an equilibrium consumers make sure that the present values of last-period monies to hold and debts are all zero. This means that in an equilibrium the effective constraints on the choice of consumption plans are:

$$\sum_{t=1}^{T} d_{Ht}p_{Ht}c_t = M_H + \sum_{t=1}^{T-1} d_{Ht+1}p_{Ht}y_{Ht} + \sum_{t=1}^{T} d_{Ht}X_{Ht}, \tag{10}$$

$$\sum_{t=1}^{T} d_{Ft}p_{Ft}c^*_t = M_F + \sum_{t=1}^{T-1} d_{Ft+1}p_{Ft}y_{Ft} + \sum_{t=1}^{T} d_{Ft}X_{Ft}. \tag{11}$$

Thus, equilibrium consumption plans are solutions to: choose $c_t \geqslant 0$, $t = 1, 2, \ldots, T$, to maximize $u_H(c_1, c_2, \ldots, c_T)$ subject to (10); and choose $c^*_t \geqslant 0$, $t = 1, 2, \ldots, T$, to maximize $u_F(c^*_1, c^*_2, \ldots, c^*_T)$ subject to (11). Due to the interest and purchasing power parity conditions $(d_{Ft}p_{Ft})/(d_{F1}p_{F1}) = (d_{Ht}p_{Ht})/(d_{H1}p_{H1})$, for $t = 1, 2, \ldots, T$, implying that consumers face the same relative prices of consumption in both

[3] This implication is a direct consequence of the assumption that one can switch costlessly from money to bonds and vice versa. With positive transaction costs associated with purchases and sales of bonds, money is used as a store of value even when interest rates are positive (see Baumol 1952 and Tobin 1956). This suggests that additional insight might be gained by extending the present model to allow for transaction costs in financial markets.

problems. Hence, every equilibrium allocation of consumption in which consumption demands are solutions to these problems is Pareto efficient. This proves the main result of this section.

It is important to observe at this stage that the implications of being in an equilibrium which have been discussed so far are independent of the exchange-rate system; they apply to systems of fixed and floating as well as managed exchange rates. Every exchange-rate system has associated with it specific financial market-clearing conditions (the clearing conditions for commodity markets, $c_{Ht} + c^*_{Ht} = y_{Ht}$ and $c_{Ft} + c^*_{Ft} = y_{Ft}$, are the same for every exchange-rate regime), which may have specific distributional implications. These implications are considered in the following sections.

IV. Floating Exchange-Rate Regime

By a floating exchange-rate regime I refer to a regime in which no government agency intervenes in the foreign exchange market for the sake of intervention. Thus, if one was to reformulate the present model to include government purchases of goods and the government was instructed to acquire foreign exchange in order to purchase foreign goods, this would introduce intervention in the foreign exchange market but *not* for the sake of intervention. However, if a government agency purchases foreign exchange, keeps it for several periods, and then resells it, we refer to this action as intervention for the sake of intervention. In the presence of this type of policy the exchange-rate regime is a managed float.

Now consider the economies described in the previous section in a floating exchange-rate regime. The governments are assumed to engage in no economic activity except for the monetary transfers X_{Ht} and X_{Ft} (which may be zero). This means that each currency has only one source of money creation—government transfers—that all debt is private debt, and that all commodity demands are private demands. Under these circumstances a floating exchange-rate regime equilibrium can be defined as follows:

DEFINITION: Prices $\{\tilde{p}_{Ht}, \tilde{p}_{Ft}\}_{t=1}^{T}$, exchange rates $\{\tilde{e}_t\}_{t=1}^{T}$, interest rates $\{\tilde{i}_{Ht}, \tilde{i}_{Ft}\}_{t=1}^{T-1}$, and the allocation $\{\tilde{c}_{Ht}, \tilde{c}_{Ft}, \tilde{c}^*_{Ht}, \tilde{c}^*_{Ft}, \tilde{M}_{Ht}, \tilde{M}_{Ft}, \tilde{M}^*_{Ht}, \tilde{M}^*_{Ft}, \tilde{B}_{Ht}, \tilde{B}_{Ft}, \tilde{B}^*_{Ht}, \tilde{B}^*_{Ft}\}_{t=1}^{T}$ constitute a floating exchange-rate regime equilibrium relative to $[u_H, u_F, M_H, M_F, \{X_{Ht}, X_{Ft}, y_{Ht}, y_{Ft}\}_{t=1}^{T}]$, if:

a) for $p_{Ht} = \tilde{p}_{Ht}, p_{Ft} = \tilde{p}_{Ft}, e_t = \tilde{e}_t, t = 1, 2, \ldots, T$, and $i_{Ht} = \tilde{i}_{Ht}, i_{Ft} = \tilde{i}_{Ft}$, $t = 1, 2, \ldots, T - 1$,

i) $\{\tilde{c}_{Ht}, \tilde{c}_{Ft}, \tilde{M}_{Ht}, \tilde{M}_{Ft}, \tilde{B}_{Ht}, \tilde{B}_{Ft}\}_{t=1}^{T}$ solves (1),

ii) $\{\tilde{c}^*_{Ht}, \tilde{c}^*_{Ft}, \tilde{M}^*_{Ht}, \tilde{M}^*_{Ft}, \tilde{B}^*_{Ht}, \tilde{B}^*_{Ft}\}_{t=1}^{T}$ solves (2),

b) $\tilde{c}_{Ht} + \tilde{c}^*_{Ht} = y_{Ht}$, $t = 1, 2, \ldots, T$,

c) $\tilde{c}_{Ft} + \tilde{c}^*_{Ft} = y_{Ft}, t = 1, 2, \ldots, T,$

d) $\tilde{M}_{Ht} + \tilde{M}^*_{Ht} = M_H + \sum_{\tau=1}^{t} X_{H\tau}, t = 1, 2, \ldots, T,$

e) $\tilde{M}_{Ft} + \tilde{M}^*_{Ft} = M_F + \sum_{\tau=1}^{t} X_{F\tau}, t = 1, 2, \ldots, T,$

f) $\tilde{B}_{Ht} + \tilde{B}^*_{Ht} = 0, t = 1, 2, \ldots, T,$

g) $\tilde{B}_{Ft} + \tilde{B}^*_{Ft} = 0, t = 1, 2, \ldots, T.$

In an equilibrium, individual optimization based on self-fulfilling expectation (condition *a*) is consistent with clearance of all markets in every period; market for home output (condition *b*), market for foreign output (condition *c*), market for home money (condition *d*), market for foreign money (condition *e*), market for home currency-denominated debt (condition *f*), and market for foreign currency-denominated debt (condition *g*).

The right-hand side of condition *d* represents the supply of home money in period *t*; it consists of the initial stock of money plus monetary injections up to (and including) period *t*. Similarly, the right-hand side of condition *e* represents the supply of foreign money to the private sector in period *t*.

The main point that I wish to make in this section is that—independent of the transfer policies (X_{Ht} and X_{Ft})—in a floating exchange-rate regime the equilibrium allocation of consumption conforms to a barter equilibrium allocation of the economies under consideration, and vice versa. That is, if we allow the economies to exchange goods over time, making all trading decisions in the beginning of the first period (as in an Arrow-Debreu model), the resulting equilibrium allocation of consumption will be an equilibrium allocation of consumption in a floating exchange-rate regime, and vice versa. This is independent of the transfer policies (or, as some would call it, of monetary policies).

Despite the fact that the equilibrium allocation of consumption is determined only by real factors (the factors which determine an exchange equilibrium), every real equilibrium has associated with it equilibrium values of monetary variables like commodity prices, exchange rates, and interest rates (cf. Kareken and Wallace, in press). The monetary variables are determined within the system.

In order to prove the main result of this section recall the formulation of the optimal consumption plan problems that was developed at the end of Section III.

Choose

$$c_t \geqslant 0, \qquad t = 1, 2, \ldots, T, \tag{12}$$

to maximize

$$u_H(c_1, c_2, \ldots, c_T)$$

subject to

$$\sum_{t=1}^{T} d_{Ht} p_{Ht} c_t = M_H + \sum_{t=1}^{T-1} d_{Ht+1} p_{Ht} y_{Ht} + \sum_{t=1}^{T} d_{Ht} X_{Ht}.$$

Choose

$$c_t^* \geq 0, \qquad t = 1, 2, \ldots, T, \tag{13}$$

to maximize

$$u_F(c_1^*, c_2^*, \ldots, c_T^*),$$

subject to

$$\sum_{t=1}^{T} d_{Ft} p_{Ft} c_t^* = M_F + \sum_{t=1}^{T-1} d_{Ft+1} p_{Ft} y_{Ft} + \sum_{t=1}^{T} d_{Ft} X_{Ft}.$$

Equilibrium consumption schedules are solutions to these problems independent of the exchange-rate regime.

It is shown in the Appendix that, in a floating exchange-rate regime equilibrium, the right-hand side of the wealth constraint in (12) is equal to $\Sigma_{t=1}^{T} d_{Ht} p_{Ht} y_{Ht}$ while the right-hand side of the wealth constraint in (13) is equal to $\Sigma_{t=1}^{T} d_{Ft} p_{Ft} y_{Ft}$. This means that in each country the present value of output equals the value of initial money holdings plus the present value of dividends plus the present value of government transfers. Hence, a country's assets reflect precisely the value of its real output. Now, since $(d_{Ft} p_{Ft})/(d_{F1} p_{F1}) = (d_{Ht} p_{Ht})/(d_{H1} p_{H1})$ (due to interest rate and purchasing power parity), an equilibrium allocation of consumption in a floating exchange-rate regime—which is a solution of (12) and (13)—coincides with an equilibrium allocation of consumption in a costless barter economy.[4] Hence, despite the fact that the monies play a real role in facilitating transactions, thus introducing potential divergences from barter equilibria, in a floating exchange-rate regime with perfect foresight no such divergences exist.

The fact that the equilibrium allocation of consumption in a floating exchange-rate regime coincides with the equilibrium allocation of consumption in a costless barter regime enables a rather simple calculation of equilibrium prices, exchange rates, and nominal interest rates. First, compute the equilibrium relative intertemporal prices in the barter regime. Let them be $q_1 = 1, q_2, q_3, \ldots, q_T$. Then we know

[4] This result is also derived in Lucas (1980*b*).

that in the floating exchange-rate regime prices and interest rates satisfy:

$$\frac{d_{Ht}p_{Ht}}{d_{Ht-1}p_{Ht-1}} = \frac{d_{Ft}p_{Ft}}{d_{Ft-1}p_{Ft-1}} = \frac{q_t}{q_{t-1}}, \qquad t = 2, 3, \ldots, T. \quad (14)$$

Now, in the last period all money is spent on goods. Therefore, last-period prices are

$$p_{HT} = \frac{\left(M_H + \sum_{\tau=1}^{T} X_{H\tau}\right)}{y_{HT}}, p_{FT} = \frac{\left(M_F + \sum_{\tau=1}^{T} X_{F\tau}\right)}{y_{FT}}.$$

Having last-period prices, we can determine prices and interest rates in period $T - 1$. From (14), using $d_{kt}/d_{kt-1} = (1 + i_{kt-1})^{-1}, k = H, F$, we have

$$(1 + i_{HT-1})p_{HT-1} = \frac{p_{HT}q_{T-1}}{q_T}, (1 + i_{FT-1})p_{FT-1} = \frac{p_{FT}q_{T-1}}{q_T}. \quad (15)$$

First consider p_{HT-1} and i_{HT-1}. Suppose $i_{HT-1} > 0$. Then all H money is spent on H goods in period $T - 1$ and

$$p_{HT-1} = \frac{\left(M_H + \sum_{\tau=1}^{T-1} X_{H\tau}\right)}{y_{HT-1}}.$$

Substituting this into (15), we calculate the interest rate i_{HT-1} which satisfies (15). If the calculated interest rate is nonnegative, we have the equilibrium price level and interest rate. If the calculated interest rate is negative, then the equilibrium interest rate is zero. Substituting $i_{HT-1} = 0$ into (15), we can calculate the equilibrium price p_{HT-1}. A similar procedure is applied to calculate i_{FT-1} and P_{FT-1}. It is now clear that we can use the same procedure to calculate p_{HT-2}, i_{HT-2}, p_{FT-2}, and i_{HT-2}. Thus, starting from last-period prices, all prices and interest rates can be recursively calculated.

Observe that prices are determined according to the quantity theory of money. When a country's money is not used as a store of value, all of it "chases" that country's goods, and the country's price level is determined according to the quantity equation with unitary velocity of circulation. When part of the money is used as a store of value, which may happen in this model only if the nominal interest rate is zero in this country, only part of the stock of money in circulation chases goods, and this part determines the price level. In fact, if the algorithm described above produces a zero interest rate in a certain period, the calculated price level of this period can be used in conjunction with output to calculate the quantity of money re-

quired for transactions. The residual quantity of money will be used as money to hold.[5]

V. Fixed Exchange-Rate Regime: One-sided Peg

In this section I consider a fixed exchange-rate regime in which the home country pegs its exchange rate to the foreign country. For this purpose the home country establishes an exchange-rate stabilizing authority (ERSA) whose task is to buy and sell foreign currency at the fixed rate of exchange e, which is assumed to remain the same in every period. The ERSA replaces the foreign exchange market. Since the home country cannot issue foreign money, its one-sided decision to peg the exchange rate makes it necessary for the ERSA to borrow abroad whenever it is asked to sell foreign currency, and I assume that whenever it is asked to buy foreign currency it lends it out. This assumption avoids unnecessary interest costs to the ERSA (these interest costs are sometimes considered as part of the costs of exchange-rate stabilization). Also, it is assumed that, whenever the ERSA buys foreign currency, it pays for it with new issues of home currency, thus increasing the quantity of home money in circulation. Whenever the ERSA sells foreign currency, it leaves in its possession the home currency that it receives against these sales, thereby reducing the quantity of home money in circulation. This mechanism of exchange-rate stabilization approximates practices that were used in many countries (central banks usually perform the tasks of the ERSA).

In what follows I wish to make two main points. One concerns the distributional implications of the exchange-rate system that is being considered (we know from Sec. II that the equilibrium allocation of consumption is efficient). It will be shown that when the ERSA is constrained to end up with no foreign debt, the resulting equilibrium allocation of consumption is the same as in a costless barter economy and, therefore, also the same as in a floating exchange-rate regime. Second, in order to meet the ERSA's intertemporal budget constraint (the constraint on last period's foreign debt), the transfer policy cannot be arbitrary. This means that in the present system the transfer policy cannot be divorced from the stabilization of the exchange rate, which may be considered as a disadvantage of this fixed exchange-rate regime.[6] Apart from these points we will have a closer look at the

[5] An explicit calculation of an equilibrium for two toy economies is presented in Helpman (1979). Helpman and Razin (1981) present a similar calculation for the infinite horizon case.

[6] A similar argument, but in a different setup, appears in Helpman and Razin (1979).

behavior of prices and money stocks and the relationship between the exchange rate and the required transfer or absorption policy.

I start by presenting the ERSA's budget constraint. In order to save space let us use the convention that variables with a subscript $t - 1$ are all zero for $t = 1$. The ERSA's foreign debt in period t is denoted by B^E_{Ft} and its addition to H money in circulation in period t is denoted by X^E_{Ht} (X^E_{Ht} may, of course, be positive or negative). The exchange rate is constant; that is, $e_t = e$.

There are two economic units which actively transact with the ERSA: the private sector of the home country and the private sector of the foreign country. The private sector of the home country has in period t an excess demand for foreign currency which equals $M_{Ft} + (1 + i_{Ft-1})B_{Ft-1} - m_{Ft-1} - B_{Ft}$; it equals total desired holdings of foreign currency, plus foreign debt repayments, minus the quantity of foreign currency transferred from the previous period (the previous period's foreign money to hold), minus foreign borrowing. This excess demand is satisfied by purchases of foreign currency from the ERSA, for which the private sector pays with home country money. The foreign private sector has an excess demand for home currency which equals $M^*_{Ht} + (1 + i_{Ht-1})B^*_{Ht-1} - m^*_{Ht-1} - B^*_{Ht}$. This excess demand is also satisfied by transacting with the ERSA. For the purchase of home currency the foreign private sector pays $[M^*_{Ht} + (1 + i_{Ht-1})B^*_{Ht-1} - m^*_{Ht-1} - B^*_{Ht}]/e$ units of foreign currency. Hence, the deficit in the balance of payments which equals the net excess demand for foreign currency with which the ERSA is faced in period t is:

$$D_{Ft} = [M_{Ft} + (1 + i_{Ft-1})B_{Ft-1} - m_{Ft-1} - B_{Ft}] - \frac{[M^*_{Ht} + (1 + i_{Ht-1})B^*_{Ht-1} - m^*_{Ht-1} - B^*_{Ht}]}{e}. \tag{16}$$

Against this excess demand for foreign currency there is an equal value excess supply of home currency which the ERSA has to absorb. Hence, $X^E_{Ht} = -eD_{Ft}$, and

$$X^E_{Ht} = -e[M_{Ft} + (1 + i_{Ft-1})B_{Ft-1} - m_{Ft-1} - B_{Ft}] + [M^*_{Ht} + (1 + i_{Ht-1})B^*_{Ht-1} - m^*_{Ht-1} - B^*_{Ht}]. \tag{17}$$

Apart from the net excess demand for foreign currency by the private sector in period t, the ERSA needs foreign currency in order to repay the debt that it acquired in period $t - 1$. Hence, the ERSA's foreign borrowing in period t equals the private sector's net excess demand for foreign currency plus its repayments of foreign debt:

$$B^E_{Ft} = [M_{Ft} + (1 + i_{Ft-1})B_{Ft-1} - m_{Ft-1} - B_{Ft}] - \frac{[M^*_{Ht} + (1 + i_{Ht-1})B^*_{Ht-1} - m^*_{Ht-1} - B^*_{Ht}]}{e} + (1 + i_{Ft-1})B^E_{Ft-1}. \tag{18}$$

The last equation is based on the assumption that the ERSA does not carry forward any foreign currency; when it has foreign currency it either sells it or lends it out. This assumption assures efficient management of the exchange-rate stabilizing operations.[7] From (16) and (18), we have:

$$D_{Ft} = B^E_{Ft} - B^E_{Ft-1} - i_{Ft-1}B^E_{Ft-1}. \tag{19}$$

The deficit in the balance of payments equals the increase in the ERSA's foreign debt minus its interest payments. This is the definition of the deficit in the balance of payments which associates the deficit with the ERSA's net sales of foreign currency to the private sectors, which also reflects the contraction of home currency in circulation due to the balance of payments. However, in this framework it might be more appropriate to define the deficit as the increase in the ERSA's foreign debt. The difference between the former and latter definitions amounts to the exclusion or inclusion, respectively, of interest payments on the ERSA's foreign debt in the definition of the deficit in the balance of payments. Clearly, if the latter definition is used, then the contraction in home money in circulation as a result of exchange-rate stabilization is not equal to the deficit in the balance of payments.

Now we are prepared to define a one-sided peg fixed exchange-rate regime equilibrium. A formal definition is provided in Helpman (1979). Here I discuss informally the difference between the one-sided peg equilibrium and the floating exchange-rate equilibrium.

In the one-sided peg the ERSA stabilizes the exchange rate at a fixed level e in all periods. Hence $e_t = e$ for $t = 1, 2, \ldots, T$. As a result, it engages in foreign borrowing (B^E_{Ft}) and monetary injections (X^E_{Ht}). Since the ERSA has to satisfy its temporal budget constraints and break even over time, all choice variables—private as well as the ERSA's—have to satisfy (17) and (18). In addition, the exchange-rate stabilizing authority has to end up with no foreign debts:

$$B^E_{FT} = 0. \tag{20}$$

Of course, commodity markets have to clear, so that conditions b and c from the definition of a floating exchange-rate equilibrium apply also to the present regime. Finally, money markets and bond markets have to clear. Since our ERSA does not operate in the domestic bond market, it does not carry forward foreign money, and it does not make foreign currency transfers, conditions e and f from the definition of a floating exchange-rate equilibrium apply also to the present regime. A difference arises in the equilibrium conditions for the home currency market and the foreign bond market, because the

[7] See, however, Helpman and Razin (1980) for a case in which, due to uncertainty, the ERSA has to carry forward some foreign currency.

ERSA borrows on the foreign bond market and it changes the supply of domestic money by means of the exchange-rate stabilization operations. The market-clearing conditions for the domestic money market and the foreign bond market read in the present regime:

$$M_{Ht} + M^*_{Ht} = M_H + \sum_{\tau=1}^{t} (X_{H\tau} + X^E_{H\tau}), \quad t = 1, 2, \ldots, T, \tag{21}$$

$$B_{Ft} + B^*_{Ft} + B^E_{Ft} = 0, \quad t = 1, 2, \ldots, T. \tag{22}$$

First let me prove that the one-sided peg fixed exchange-rate regime equilibrium allocation of consumption is the same as that in a barter economy. For this purpose recall the formulation of the optimal consumption plans that was developed at the end of Section III, and let us use formulations (12) and (13) for present purposes. It is sufficient to show that in the home country's problem (12), the right-hand side of the wealth constraint equals the present value of output in equilibrium.

In a fixed exchange-rate regime the interest rate parity condition (3) reads $i_{Ht} = i_{Ft} = i_t$, $t = 1, 2, \ldots, T$, which implies equality of discount factors; that is, $d_{Ht} = d_{Ft} = d_t$, $t = 1, 2, \ldots, T$. Hence, the equality between the right-hand side of the wealth constraint in (12) and the present value of output translates in the present context into:

$$M_H + \sum_{t=1}^{T-1} d_{t+1} p_{Ht} y_{Ht} + \sum_{t=1}^{T} d_t X_{Ht} = \sum_{t=1}^{T} d_t p_{Ht} y_{Ht},$$

which is equivalent to

$$\sum_{t=1}^{T-1} (d_{t+1} - d_t) p_{Ht} y_{Ht} - d_T p_{HT} y_{HT} = -M_H - \sum_{t=1}^{T} d_t X_{Ht}. \tag{23}$$

Now if $i_t > 0$ (i.e., $d_{t+1} < d_t$), money is not used as a store of value and all home money is spent on home goods, implying:

$$(d_{t+1} - d_t) p_{Ht} y_{Ht} = (d_{t+1} - d_t)\left[M_H + \sum_{\tau=1}^{t} (X_{H\tau} + X^E_{H\tau})\right]. \tag{24}$$

If $i_t = 0$ (i.e., $d_{t+1} = d_t$), then (24) holds trivially. Hence, (24) holds for $t = 1, 2, \ldots, T - 1$. In the last period all money is spent on goods; therefore $p_{HT} y_{HT} = M_H + \Sigma_{\tau=1}^{T} (X_{H\tau} + X^E_{H\tau})$. Combining the last result with (24) and applying the method employed in the Appendix, one gets:[8]

$$\sum_{t=1}^{T-1} (d_{t+1} - d_t) p_{Ht} y_{Ht} - d_T p_{HT} y_{HT} = -M_H - \sum_{t=1}^{T} d_{Ht} (X_{Ht} + X^E_{Ht}).$$

[8] For details see Helpman (1979).

By comparing the last result with (23) it is clear that in order to prove (23) it remains to show that $\Sigma_{t=1}^{T} d_{Ht} X_{Ht}^{E} = 0$. From (17) and (18), using the interest rate parity condition $i_{Ht} = i_{Ft} = i_t$, we have:

$$X_{Ht}^{E} = e[(1 + i_{t-1})B_{Ft-1}^{E} - B_{Ft}^{E}].$$

Hence, using the convention $B_{F0}^{E} = 0$ and the fact that $d_t(1 + i_{t-1}) = d_{t-1}$, we calculate

$$\sum_{t=1}^{T} d_t X_{Ht}^{E} = e \sum_{t=1}^{T} d_t[(1 + i_{t-1})B_{Ft-1}^{E} - B_{Ft}^{E}]$$

$$= e\left(\sum_{t=2}^{T} d_{t-1} B_{Ft-1}^{E} - \sum_{t=1}^{T} d_t B_{Ft}^{E}\right) = -e d_T B_{FT}^{E} = 0,$$

which completes the proof (the last equality results from [20]).

I have shown that the equilibrium allocation of consumption in the present exchange-rate regime is the same as in a costless barter economy and, therefore, also the same as in an equilibrium of a floating exchange-rate regime. This means that the welfare level achieved by each country is the same in each (equilibrium) of these three alternative modes of organization of the world economy. What I wish to argue now is that the country that chooses to stabilize the exchange rate in the one-sided peg fixed exchange-rate regime cannot pursue an independent monetary (transfer) policy. In particular, for given transfer policies in the two countries, there is exactly *one* level at which it can stabilize the exchange rate and end up with no foreign debt. Put differently, if the country chooses to stabilize the exchange rate at a particular level, the requirement that it ends up with no foreign debt imposes on it a particular transfer policy; the present value of its transfers is uniquely determined, although their distribution over time can be freely chosen. Changes in the pattern of transfers change the pattern of deficits in the balance of payments but not the equilibrium pattern of consumption, provided the present value of transfers is maintained at the equilibrium level.

In order to prove my arguments, observe that the purchasing power parity condition $p_{Ht} = ep_{Ft}$ and (23) imply:

$$M_H + \sum_{t=1}^{T} d_t X_{Ht} = e\left[d_T p_{FT} y_{HT} + \sum_{t=1}^{T-1} (d_t - d_{t+1}) p_{Ft} y_{Ht}\right]. \tag{25}$$

Since in the last period foreign money is spent on foreign goods,

$$p_{FT} = \frac{\left(M_F + \sum_{t=1}^{T} X_{Ft}\right)}{y_{FT}}. \tag{26}$$

If $i_t > 0, t \in \{1, 2, \ldots, T-1\}$, then in period t all foreign money is spent on foreign goods. If $i_t = 0$, then $d_t = d_{t+1}$. Therefore,

$$(d_t - d_{t+1})p_{Ft} = (d_t - d_{t+1}) \frac{\left(M_F + \sum_{\tau=1}^{t} X_{F\tau}\right)}{y_{F\tau}}, \quad t = 1, 2, \ldots, T-1. \tag{26a}$$

In addition,

$$\frac{d_t p_{Ft}}{d_{t-1} p_{Ft-1}} = \frac{q_t}{q_{t-1}}, \quad t = 2, 3, \ldots, T, q_1 = 1, \tag{26b}$$

where q_t is the equilibrium marginal rate of substitution between consumption in period t and consumption in period 1. From (26) foreign prices and the discount factors are uniquely determined. They are, in fact, identical to the equilibrium foreign prices and foreign discount factors in a floating exchange-rate regime because the equilibrium marginal rates of substitution q_t are the same in both regimes. Hence, the equilibrium value of the expression in the brackets on the right-hand side of (25)—which is positive—is independent of the home country's transfer policy and exchange-rate level. Therefore (25) implies a one-to-one relationship between the exchange-rate level and the present value of home money transfers.

VI. Fixed Exchange-Rate Regime: Cooperative Peg

The last exchange-rate regime that I wish to consider is one in which both countries cooperate in the stabilization of the exchange rate. For this purpose the countries establish an ERSA which buys and sells currencies at a fixed rate of exchange, e units of home currency per unit of foreign currency. The cooperation consists of each country's supplying the ERSA with its own currency whenever the ERSA is asked to sell it and is short on it. The ERSA does not engage in borrowing and lending. When foreign currency is sold to the ERSA, this transaction reduces the quantity of foreign currency in circulation. It also increases the quantity of home currency in circulation, with the increase equal in value to the reduced amount of foreign currency in circulation, since sales of foreign currency mean purchases of home currency. The opposite effect occurs when home currency is sold to the ERSA. This means that the operations of the ERSA do not change the aggregate value of monies in circulation; they only change their composition. However, transfer policies affect the aggregate quantity of money in circulation. The present regime is in line with recent modelings of fixed exchange-rate regimes in that

the stabilization of the exchange rate per se does not affect the aggregate supply of money in the world economy.[9]

In the cooperative fixed exchange-rate regime the changes in money supply caused by the ERSA are (using the interest parity condition and the convention that variables with subscripts $t - 1$ are zero for $t = 1$):

$$X_{FT}^{E} = [M_{Ft} + (1 + i_{t-1})B_{Ft-1} - m_{Ft-1} - B_{Ft}] - \frac{[M_{Ht}^{*} + (1 + i_{t-1})B_{Ht-1}^{*} - m_{Ht-1}^{*} - B_{Ht}^{*}]}{e}, \tag{27}$$

$$X_{Ht}^{E} = [M_{Ht}^{*} + (1 + i_{t-1})B_{Ht-1}^{*} - m_{Ht-1}^{*} - B_{Ht}^{*}] - e[M_{Ft} + (1 + i_{t-1})B_{Ft-1} - m_{Ft-1} - B_{Ft}]. \tag{28}$$

Thus,

$$X_{Ht}^{E} + eX_{Ft}^{E} = 0, \qquad t = 1, 2, \ldots, T. \tag{29}$$

The definition of an equilibrium should by now be obvious and I abstain from stating it.

Using the equilibrium and parity conditions, one derives—by means of the method employed in the previous section—the following budget constraints on consumption vectors:

$$\sum_{t=1}^{T} d_t p_{Ht} c_t = \sum_{t=1}^{T} d_t p_{Ht} y_{Ht} - \sum_{t=1}^{T} d_t X_{Ht}^{E}, \tag{30}$$

$$\sum_{t=1}^{T} d_t p_{Ht} c_t^{*} = \sum_{t=1}^{T} d_t p_{Ht} y_{Ft} - e \sum_{t=1}^{T} d_t X_{Ft}^{E}. \tag{30a}$$

Due to (29), (30) implies that the equilibrium allocation of consumption in the cooperative fixed exchange-rate regime coincides with the equilibrium allocation of consumption in a barter economy if and only if the present value of monetary injections of currency H (or contractions of currency F) caused by the operations of the ERSA is zero; that is, $\Sigma_{t=1}^{T} d_t X_{Ht}^{E} = 0$. I have shown in the previous section that in the one-sided fixed exchange-rate regime $\Sigma_{t=1}^{T} d_t X_{Ht}^{E}$ is necessarily zero in an equilibrium. Here this is not the case; in an equilibrium of the present regime the present value of home currency injections

[9] If in a cooperative peg only one country sterilizes changes in its money supply that result from balance-of-payments adjustments, then the behavior of money supplies in this regime will resemble the behavior of money supplies in a one-sided peg regime. It is interesting to note that the prediction of the one-sided peg regime about the behavior of the world price level is consistent with the empirical findings in Genberg and Swoboda (1977) (i.e., that the money supply of the passive country, the United States, explains the world's dollar price level). Since they took the standard view of a fixed exchange rate (i.e., a cooperative peg), they explained their findings with an argument about sterilization on the part of the United States, an argument that they did not test.

through the balance of payments need not be zero.[10] If it is positive, the home country's present value of consumption is less than the present value of its output, and the foreign country's present value of consumption exceeds its present value of output. If it is negative, the opposite holds. In the former case it is as if the home country is making a transfer to the foreign country, while in the latter case it is as if the foreign country is making a transfer to the home country. Since in either case the allocation is efficient, then—to the extent that the transferer loses in welfare terms (which is not always the case)—in the former case the home country is worse off and the foreign country better off than in a floating exchange-rate regime (or a one-sided peg), while in the latter case the opposite holds. Since the outcome depends on the level of the exchange rate and on each country's transfer policy, countries have conflicting interests in the determination of the exchange-rate level and monetary policies.

VII. Concluding Remarks

I have presented in this study an evaluation of exchange-rate regimes for a world economy in which there exists a single friction; that is, there exists a payments period and monies are required for market transactions. It has been shown that this friction has certain distributional implications in each exchange-rate regime, but it does not introduce inefficiencies into the economic system. All results are easily extended to a many-country framework. In a way, these results provide a benchmark for the evaluation of more realistic models and their implications about exchange-rate regimes.

There are various rigidities and imperfections (e.g., wage contracts, market imperfection) whose existence may have a bearing on the relative desirability of alternative exchange-rate regimes. Thus, it is possible that a particular exchange-rate regime performs better than others under certain types of imperfections. A characterization of imperfections under which each exchange-rate regime performs best will be valuable. Some work on the effects of uncertainty and imperfect capital markets is reported in Helpman and Razin (1980) and Lucas (1980*b*). Lucas provides conditions under which a floating exchange-rate regime is equivalent (from the allocational viewpoint) to a fixed exchange-rate regime, despite the existence of uncertainty and incomplete capital markets, while Helpman and Razin provide conditions under which a floating exchange-rate regime performs better than a one-sided peg.

There is also another aspect of exchange-rate regimes which has

[10] An example showing this point is presented in Helpman (1979).

received no attention in a welfare context but which is of major importance, and this is the strategic behavior of governments. I have shown that the equilibrium allocation of consumption is efficient in every exchange-rate regime and that, in a floating exchange-rate regime as well as in a one-sided peg regime, it corresponds to the equilibrium allocation of consumption in a barter economy. A crucial assumption on which these results are based is that expectations are self-fulfilling; that is, there are no surprises. In particular, the transfer policies are announced at the very beginning, and each government honors its declared policy. What happens if in the course of time one of the governments deviates from its announced policy? This type of surprise destroys both results. This can be seen as follows. Suppose that at some point in time the foreign country is a net creditor in terms of foreign currency and it has zero net debt in terms of home currency. Then, by reducing the quantity of foreign money it increases via exchange-rate changes the real value of payments that it will receive for its outstanding credit.[11] Thus, governments have an incentive to generate monetary surprises. But if they do it, one cannot analyze the consumer problem the way it was done in this paper, because the consumer will learn about the possible surprises and will accordingly adjust his decisions. Other governments will also adjust their policy. Hence, a game-theoretic framework, with an appropriate reformulation of individual problems, is required in order to analyze these issues properly.

The incentive for strategic behavior is even more apparent in the cooperative-peg regime. Recall that in this regime the exchange-rate level and the preannounced monetary policies determine real transfers among countries. There are, therefore, conflicting interests in the choice of an exchange-rate level and monetary (transfer) policies. In particular, once an exchange-rate level has been chosen, then in the course of time countries may try to secure for themselves higher welfare levels by departing from the preannounced monetary policies. Such unexpected departures generate allocational inefficiencies. Again, in order to analyze this problem properly, one needs an appropriate game-theoretic framework.

The need for an extension of the present analysis in order to take account of strategic behavior is clear from the context of this paper. However, such an analysis may also help to explain changes in exchange-rate regimes that took place in the world economy, such as the collapse of the Bretton Woods system.

[11] A numerical example showing this effect is presented in Helpman (1979). Similar effects exist in the one-sided peg regime.

Appendix

Here I prove that in a floating exchange-rate regime equilibrium the right-hand side of the wealth constraint in (12) equals $\Sigma_{t=1}^{T} d_{Ht} p_{Ht} y_{Ht}$. A similar method can be used to show that the right-hand side of the wealth constraint in (13) equals $\Sigma_{t=1}^{T} d_{Ft} p_{Ft} y_{Ft}$. The proof proceeds as follows:

$$M_H + \sum_{t=1}^{T-1} d_{Ht+1} p_{Ht} y_{Ht} + \sum_{t=1}^{T} d_{Ht} X_{Ht} = \sum_{t=1}^{T} d_{Ht} p_{Ht} y_{Ht} + \left(M_H + \sum_{t=1}^{T} d_{Ht} X_t\right) + \sum_{t=1}^{T-1} (d_{Ht+1} - d_{Ht}) p_{Ht} y_{Ht} - d_{HT} p_{HT} y_{HT}. \tag{A1}$$

However, I have shown in Section III that for $i_{Ht} > 0$ (or $d_{Ht+1} < d_{Ht}$) no one uses domestic money as a store of value, and that in the last period all money is spent. These results, together with the equilibrium conditions b and d in the definition of a floating exchange-rate regime equilibrium, imply

$$(d_{Ht+1} - d_{Ht}) p_{Ht} y_{Ht} = (d_{Ht+1} - d_{Ht})\left(M_H + \sum_{\tau=1}^{t} X_{H\tau}\right), t = 1, 2, \ldots, T-1,$$

$$p_{HT} y_{HT} = M_H + \sum_{\tau=1}^{T} X_{H\tau}.$$

This is so, since when domestic money is not used as a store of value in period t, all of it is spent on goods, implying $p_{Ht}\tilde{c}_{Ht} = \tilde{M}_{Ht}$ and $p_{Ht}\tilde{c}^*_{Ht} = \tilde{M}^*_{Ht}$. Using condition b this implies that $p_{Ht} y_{Ht} = \tilde{M}_{Ht} + \tilde{M}^*_{Ht}$, which yields $p_{Ht} y_{Ht} = M_H + \Sigma_{\tau=1}^{t} X_{H\tau}$ when combined with d. Thus, the above relations hold for $i_{Ht} > 0$, but they hold also trivially for $i_{Ht} = 0$ since $i_{Ht} = 0$ if and only if $d_{Ht+1} = d_{Ht}$. Using these results, we obtain:

$$\sum_{t=1}^{T-1} (d_{Ht+1} - d_{Ht}) p_{Ht} y_{Ht} - d_{HT} p_{HT} y_{HT} = \sum_{t=1}^{T-1} (d_{Ht+1} - d_{Ht})\left(M_H + \sum_{\tau=1}^{t} X_{H\tau}\right) - d_{HT}\left(M_H + \sum_{\tau=1}^{T} X_{H\tau}\right) = -M_H - \sum_{t=1}^{T} d_{Ht} X_{Ht}. \tag{A2}$$

The last equality is obtained from the fact that

$$\sum_{t=1}^{T-1} d_{Ht+1} - \sum_{t=1}^{T-1} d_{Ht} - d_{HT} = \sum_{t=2}^{T-1} d_{Ht} - \sum_{t=1}^{T-1} d_{Ht} = -d_{H1} = -1,$$

and

$$\sum_{t=1}^{T-1} d_{Ht+1} \sum_{\tau=1}^{t} X_{H\tau} - \sum_{t=1}^{T-1} d_{Ht} \sum_{\tau=1}^{t} X_{H\tau} - d_{HT} \sum_{\tau=1}^{T} X_{H\tau}$$
$$= \sum_{t=1}^{T-2} d_{Ht+1} \sum_{\tau=1}^{t} X_{H\tau} - X_{H1} - \sum_{t=2}^{T-1} d_{Ht} \sum_{\tau=1}^{t} X_{H\tau} - d_{HT} X_{HT}$$

$$= \sum_{t=2}^{T-1} d_{Ht} \sum_{\tau=1}^{T-1} X_{H\tau} - \sum_{t=2}^{T-1} d_{Ht} \sum_{\tau=1}^{t} X_{H\tau} - X_{H1} - d_{HT} X_{HT} = - \sum_{t=1}^{T} d_{Ht} X_{Ht}.$$

Combining (A1) with (A2), we obtain

$$\dot{M}_H + \sum_{t=1}^{T-1} d_{Ht+1} p_{Ht} y_{Ht} + \sum_{t=1}^{T} d_{Ht} X_{Ht} = \sum_{t=1}^{T} d_{Ht} p_{Ht} y_{Ht}.$$

This completes the proof for the home country. A similar proof applies to the foreign country.

References

Baumol, William J. "The Transactions Demand for Cash: An Inventory Theoretic Approach." *Q.J.E.* 66 (November 1952): 545–56.

Clower, Robert W. "A Reconsideration of the Microfoundations of Monetary Theory." *Western Econ. J.* 6 (December 1967): 1–8.

Flanders, M. June, and Helpman, Elhanan. "On Exchange Rate Policies for a Small Country." *Econ. J.* 88 (March 1978): 44–58.

Friedman, Milton. "The Case for Flexible Exchange Rates." In *Essays in Positive Economics.* Chicago: Univ. Chicago Press, 1953.

Genberg, Hans, and Swoboda, Alexander K. "Worldwide Inflation under the Dollar Standard." Discussion Paper no. 12, Grad. Inst. Internat. Studies, Geneva. Revised version, January 1977.

Helpman, Elhanan. "An Exploration in the Theory of Exchange Rate Regimes." Working Paper no. 79-2, Univ. Rochester, Dept. Econ., January 1979.

Helpman, Elhanan, and Razin, Assaf. "Towards a Consistent Comparison of Alternative Exchange Rate Systems." *Canadian J. Econ.* 12 (August 1979): 394–409.

———. "A Comparison of Exchange Rate Regimes in the Presence of Imperfect Capital Markets." Seminar Paper no. 156, Univ. Stockholm, Inst. Internat. Econ. Studies, October 1980.

———. "Dynamics of a Floating Exchange Rate Regime." Working Paper no. 21-81, Tel-Aviv Univ., Foerder Inst. Econ. Res., May 1981.

Hicks, John R. "The Two Triads." Lectures 1 and 2 in *Critical Essays in Monetary Theory.* Oxford: Clarendon, 1967.

Kareken, John H., and Wallace, Neil. "Samuelson's Consumption-Loan Model with Country Specific Fiat Monies." *Q.J.E.* (in press).

Lucas, Robert E., Jr. "Equilibrium in a Pure Currency Economy." *Econ. Inquiry* 18 (April 1980): 203–20. (*a*)

———. "Notes on the Efficiency of Exchange Rate Regimes." Private notes, March 1980. (*b*)

Mundell, Robert A. *International Economics.* New York: Macmillan, 1968.

Stockman, Alan C. "A Theory of Exchange Rate Determination." *J.P.E.* 88, no. 4 (August 1980): 673–98.

Tobin, James. "The Interest-Elasticity of Transactions Demand for Cash." *Rev. Econ. and Statis.* 38 (August 1956): 241–47.

Subsidies to New Energy Sources: Do They Add to Energy Stocks?

William J. Baumol

Princeton University and New York University

Edward N. Wolff

New York University

Energy-production subsidies are, paradoxically, shown to be likely to increase U.S. dependence on imported oil. Standard studies of net energy yields are shown to be seriously biased upward for two reasons. First, many omit indirect energy inputs, which, our input-output calculation shows, probably causes large errors. Second, those studies all omit the energy opportunity costs of nonenergy inputs (e.g., the fuel substituted elsewhere for the labor used to produce energy). We prove that, absent externalities, any fuel-output subsidy which causes an otherwise unprofitable expansion must yield an incremental fuel output smaller than the increment in energy input plus the energy opportunity costs of other inputs.

The policy measures under consideration to decrease our dependence on imported oil now seem to include two leading candidates: (i) taxes or tariffs to discourage consumption and increase the profitability of alternative energy sources and (ii) a program of subsidies to innovative energy sources. Politicians have generally embraced the latter as the one means of salvation from an alternative

We are sincerely grateful to the Sloan Foundation, whose grant greatly facilitated completion of this study. Sue Anne Batey Blackman provided invaluable research on net-energy study techniques in practice and has given us very helpful comments. We are also grateful to James Sweeney of Stanford University for valuable comments on an earlier draft which helped us to avoid some misleading statements of our results.

[*Journal of Political Economy*, 1981, vol 89, no. 5]
© 1981 by The University of Chicago. 0022-3808/81/8905-0001$01 50

that appears to them to be politically suicidal, while economists have almost uniformly favored the former, arguing that the program of special subsidies is an invitation to inefficiency.

In this paper we will come to a similar but considerably stronger conclusion. We will argue that the difference between the two approaches is potentially more significant and serious than even most economists had imagined. Our main conclusion can best be stated in the form of an example. Both programs we are considering can be used to stimulate the production of oil from sources such as shale. Any expansion in this activity stimulated by a *tariff* on imported oil can be expected to have a positive net energy yield and, in particular, a positive net yield of oil. However, an expansion in the same activity induced by a program of *subsidies* is likely to have a negative net yield of petroleum (!) if other outputs in the economy remain unchanged. Of course, in practice the subsidy is likely to cause some reallocation of resources that reduces at least some other (energy-using) economic activities. The final result may well be an overall increase in the nation's stock of energy. However, we will show, under very plausible assumptions, that the increase in oil stocks could always have been made greater still if exactly the same reduction in other energy-using activities had occurred by itself, with no expansion in oil shale processing. The implications for current policy are, clearly, more than a little disquieting.

This result actually has a straightforward explanation, as we will see. Basically, the extreme difference between the efficiency of the general tariff and that of the targeted subsidy lies in the fact that the latter does not lead to the appropriate substitutions of energy and other inputs throughout the economy that should be part of a program to increase oil stocks. Consequently, the subsidy induces the employment in oil shale processing of inputs which could have been substituted elsewhere in the economy for larger quantities of oil than they can yield through oil shale conversion.

Several comments are required before we turn to the body of our analysis. First, the basic premise on which the entire discussion proceeds is that vulnerability of the nation to blackmail by foreign governments is a true social cost and can legitimately be treated as a detrimental externality of any increase in oil consumption. This and self-defense against economic exploitation by the cartels constitute the justification of a tariff on oil imports. In the absence of these two problems it could be argued that the rising cost of petroleum is an ordinary resource-depletion problem which can, perhaps, be handled reasonably well (distributive considerations aside) by the market, uninfluenced by any particular government intervention.

Second, it should be obvious that government programs other than the two we are considering are also possible. For example, subsidies to new energy sources can be accompanied by a set of taxes on use of oil which together will constitute a reasonably efficient program. This is rather analogous to the substitution of a combined program of taxes and subsidies for a system of effluent charges.[1] However, because of the difficulties of designing and administering an ideal mixed program, together with the fact that it is (apparently) not being considered by decision makers anywhere, we have chosen not to deal explicitly with such alternative measures.

Finally, it must be emphasized that while we believe that a tariff or some close substitute is the best way to encourage investment in solar energy, biomass, and other innovative energy sources, and that direct subsidies are likely actually to yield net energy losses in the sense indicated, we will see that there can be exceptions. Where substantial outlays on research and development must precede a substantial increase in such an energy-producing activity, and if they introduce significant free-rider problems, then for obvious reasons the market will not allocate a sufficiently large quantity of resources for the purpose. In such a case a subsidy may be entirely justified. But our analysis suggests that such exceptions should be permitted only on the basis of extremely strong evidence, in light of the danger that such a subsidy will actually exacerbate the nation's energy problem.

In this paper we will also show that the standard net-energy studies are characterized by a bias toward overestimation of the net yields of the processes they examine, only part of which has previously been recognized. This bias is not a matter of overenthusiasm of investigators, which leads them inadvertently or deliberately to distort their results, but is the consequence of several technical shortcomings in their study procedures. One of their difficulties is that, while net-energy studies commonly take into account the direct fuel inputs used in the processes which they examine, many of them consider only part of the remainder of the multiplier-like series—the materials and machines needed to make those materials and machines, and so on—all of which, obviously, have their own energy requirements. We will provide empirical evidence which indicates that the omissions are sometimes far from negligible. Moreover, even when the total energy requirements of a new process are "correctly" calculated using input-output techniques, only the *average* energy requirements per unit of

[1] We are indebted to James Sweeney for drawing our attention to this sort of mixed policy. Note, however, that a mixed program is not likely to be a perfect replacement for a tariff on oil imports. Thus, as is well known, in environmental programs partial substitution of subsidies for effluent charges is likely to lead to inferior results.

energy output will thereby be determined. The marginal energy requirements are likely to be higher than the average requirements, and therefore the marginal net energy yield may be considerably lower than the calculated average energy yield.

But this is not their only omission. The nonenergy resources that go into an energy-production process have an energy opportunity cost, as we have already noted, for they would instead have been used, at least in part, as substitutes for energy in other sectors of the economy. This energy opportunity cost is generally left out of net-energy studies, though it is in fact a very real concomitant of the production process. As a result of all of the biases that have just been mentioned, a process which appears to pass the net-energy test by a narrow margin is nevertheless likely to fail it in fact.

While the indirect energy costs and the energy opportunity costs of a production process may prove very difficult to measure directly, we will show here that there is (at least at the margin) a duality relation between profitability and net energy yield which permits the former to stand as a proxy for the latter. We will prove that any energy-production process which is unprofitable can from this fact alone be considered a potential net energy loser in the sense indicated earlier. Consequently, we can no longer presume that energy-losing projects are apt to be rare and pathological exceptions. Rather, we are forced to regard *any* innovative energy project which can be carried out only with the aid of government subsidies as a proper subject of suspicion.

I. The Energy Opportunity Cost of Nonenergy Inputs

The first issue we raise is the failure of many net-energy studies to carry their calculation of the energy used in the production of energy a sufficient number of "rounds" into the past, to determine the energy content of the inputs used to make the inputs, and so on, used in the energy-production process. Many of the net-energy studies employ a procedure called "process analysis" which involves explicit engineering descriptions of all the steps involved in a manufacturing process, and it is easy to see why it is impractical for this method to encompass many earlier rounds. Although this problem has previously been recognized in the literature, the magnitude of the resultant bias seems not to have been investigated. We therefore carried out some calculations to evaluate the magnitude of the error,[2] using 1972 input-output

[2] See, e.g., Berry 1978–79. It is noteworthy that some of these studies are careful to warn the readers of this source of bias (see, e.g., Perry et al. 1977; Berry 1978–79; Fels and Williams 1978), while others (e.g., Baron 1976, 1978; Hunt 1976) do not mention the subject.

data provided by the Bureau of Economic Analysis[3] and 1967 energy coefficients provided by the Center for Advanced Computation (1975). Letting the matrix A represent the Leontief interindustry coefficient matrix and D the vector of direct primary energy consumed per unit of sectoral output,[4] the total energy consumed per unit of sectoral output E is given, for an average sector, by $E = D(I - A)^{-1}$.

In some process-analysis studies, only the direct energy cost per unit of output of a new energy-producing process is considered, only taking account of the vector D. In others, the energy cost includes the "first-round" energy content of the new process and is evaluated, in effect yielding the sum $(D + DA)$. Reportedly, only the most ambitious process analyses include the "second-round" content, evaluating the energy cost $(D + DA + DA^2)$. Thus even these studies omit the portion $(DA^3 + DA^4 + \ldots)$ of the cost of a new energy process, and so they overestimate its net energy yield correspondingly.

We carried out calculations to estimate the magnitude of the errors that can result.[5] We first looked at the discrepancy between the direct

[3] For a description of the data, see Ritz (1979) and Ritz, Roberts, and Young (1979).

[4] The direct energy requirements for each sector are valued in British thermal units and were divided into four sources in the original data: (1) coal, (2) petroleum products, (3) natural gas, and (4) electricity. This breakdown created an immediate problem for the calculation of total energy requirements, for electricity is not a primary energy source but a manufactured form of energy. It must be produced from coal, petroleum, natural gas, uranium, running water, geothermal energy, solar energy, or some other natural source. Thus there is double counting in the original energy matrix which includes both the oil, coal, and natural gas used to produce electricity and the electricity itself. Therefore, we first assigned the primary-energy sources used to produce electricity to the electricity-using sectors, since we still want to classify the electricity as a direct energy input. Thus we pretended, in effect, that each industry produces its own electric power (as some, indeed, do). In the 85-sector scheme, there are two sectors that sell electricity—utilities (sector 68) and state and local enterprises (sector 79), which includes public utilities. Our first step was, then, to distribute the coal, petroleum, and natural gas used in the two utility sectors to the nonutility sectors in proportion to the *dollar purchases* of the nonutility sectors from the utility sectors. The next step in the assignment of electricity generation among its users was to distribute the intermediate inputs of the two utility sectors to the nonutility sectors in proportion to the dollar purchases by the latter from the former. The last step was to enter zeros into the cells recording the purchases from the utility sectors by the nonutility sectors.

[5] We carried out two more modifications in our estimation procedure. The first was to take into account the energy content of the fixed capital stock. (The A matrix consists only of intermediate flows.) Since a piece of capital is used to produce more than one unit of output, the appropriate addition is the total energy content of the capital stock of each sector per unit (or, in our case, per dollar) of output. To estimate this, we added a row of depreciation coefficients (depreciation per dollar of output) to the A matrix. Moreover, we evaluated the energy content of the depreciated portion of the capital stock per dollar of output not at its historical value but at its replacement value. To do this, we assumed that new capital stock by component was added in each sector in proportion to the sectoral components of the gross private capital formation vector in final demand. A new coefficient column was therefore added to the A matrix corre-

energy requirements and the total energy requirements in the manufacturing sectors in the input-output table. The (weighted) average ratio D_i/E_i was 0.36 (where the subscript i refers to sector i), meaning that such a calculation, on the average, omits some 64 percent of the total energy cost. The individual ratios ranged from 0.09 in the apparel industry to 0.62 in chemicals. Next, we looked at the ratio of the direct plus first-round energy requirements ($D + DA$) to the total energy requirements, which averaged 0.66 and ranged from 0.30 in apparel to 0.83 in chemicals. Finally, the ratio of direct plus first-round plus second-round energy requirements ($D + DA + DA^2$) to total energy requirements averaged 0.78 and began from as low a figure as 0.53. Even using the direct plus first-round plus second-round estimates of energy requirements, we find an average error of the order of 28 percent, which can severely bias standard estimates of net energy yield. This means that if a process-analysis calculation indicates that a given process will offer a positive net energy yield as high as, say, 20 percent, it is in fact likely to be a net energy loser.

But the omission of some indirect energy uses by net-energy studies is not the only significant source of overoptimism in their estimates of the net yield of new energy sources, for the calculation still does not take into account the potential energy-saving powers of the labor, the metals, and the other primary resources (aside from energy) used directly and indirectly in the production of energy. We will provide a formal discussion of this point, whose implication will be that *any* unprofitable energy source is a proper candidate for suspicion that it will be a net energy loser, and in later sections still stronger propositions will be offered. But first a simple example may help to indicate the nature of the issue.

Suppose that an economy has two outputs, manufactured energy and bread, and two primary resources, energy and labor; that bread can be produced by either of two processes, one labor intensive and one energy intensive; and that both processes are in fact currently in use. Then, other things being equal, if employment is constant, every labor hour used in the manufacture of energy increases the amount of bread that must be produced by the energy-intensive rather than the labor-intensive process. The corresponding increase in energy use

sponding to the depreciation row which was equal to the percentage distribution of gross private capital formation over its components. The second adjustment was to add some measure of the energy content of the (intermediate) imports used in production. Since we are not really concerned with the foreign energy content of the imports but of the domestic energy opportunity cost, the relevant measure is the energy content of the exports traded in exchange for these imports. To calculate this, we added a vector of export coefficients corresponding to the import coefficient row. This procedure implicitly values the energy content of the imports at the energy content of a corresponding dollar's worth of exports.

is the energy-opportunity cost of the hour of labor.[6] Obviously, this should be included in a full calculation of net energy. If the manufacture of 1,000 Btu of energy uses up 900 Btu directly and indirectly and requires labor time which would instead have been employed as a substitute for another 200 Btu, the process in question clearly depletes society's stock of energy on balance. A calculation which omits this opportunity cost is simply incorrect and misleading.

II. Profit as an Indicator of Energy Opportunity Cost

Direct estimates of the magnitudes of such opportunity costs may not be easy to obtain, but we will show that the profits of the energy manufacturing firm can tell us a great deal about the magnitude of the energy opportunity cost of its operations. Let the economy's production function for manufactured energy satisfy

$$F(y_1, \ldots, y_m, r_{11}, \ldots, r_{1n}, r_1, \ldots, r_n) = 0, \tag{1}$$

where y_1 is the quantity of energy output, r_{1j} is the quantity of primary input j used in energy production, and r_j is the quantity of primary resource j used as input, directly and/or indirectly in all other outputs. Let r_{11} and r_1 represent the quantities of energy input, for the moment taking all energy to be homogeneous.[7]

[6] The reality and significance of the indirect energy cost of energy production and of the energy opportunity cost of its other inputs is underscored by recent newspaper accounts. A fall in demand for nuclear electricity has led to a decrease in production of enriched uranium sufficient, according to the editor of *Electrical World* magazine, to reduce electricity use in the enrichment process by an amount equal to 1 percent of U.S. electricity consumption (*New York Times* 1980). And a page-1 story (Shabecoff 1980) indicates that rising energy prices are leading to a long-term shift toward the use of more labor in place of energy (e.g., use of additional workers to insulate homes and factories and even the hiring of more cowboys to drive cattle overland rather than transport them by truck), and it reports an estimate by Dale Jorgenson and E. A Hudson which suggests that a doubling of energy prices can increase employment by 5 percent. Although some may think that these estimates exaggerate the magnitudes of the indirect energy costs and the energy opportunity costs, they clearly show that these are real and tangible phenomena which can be expected to become increasingly obvious if energy prices continue to rise.

[7] Here, the y_i's refer to *net* output—that is, final output—instead of total output Thus the vector of final demand is

$$y = \begin{pmatrix} y_1 \\ \cdot \\ \cdot \\ \cdot \\ y_m \end{pmatrix}.$$

An intermediate output will, essentially, be valued in terms of its primary-resource content. More specifically, using a standard input-output structure, define

Then the production of energy will be profitable if[8]

$$p_1 y_1 \geqslant \Sigma p_j r_{1j}, \tag{2}$$

where p_1 is the price of both the energy output and the energy input

---

$$q = \begin{pmatrix} q_{11} & \cdots & q_{1m} \\ \cdot & & \cdot \\ \cdot & & \cdot \\ \cdot & & \cdot \\ q_{n1} & \cdots & q_{nm} \end{pmatrix}$$

to be the n by m coefficient matrix, showing the amount of each primary input 1 through n required per unit of output for each of the m sectors. We will let the first row of q contain the direct energy inputs in each of the sectors per unit of output. Therefore,

$$q = \begin{pmatrix} d_1 & \ldots & d_m \\ q_{21} & & q_{2m} \\ q_{n1} & \cdots & q_{nm} \end{pmatrix}.$$

The sum of the direct and indirect primary inputs required per unit of final demand is then given by matrix s:

$$s = q(I - A)^{-1} = \begin{pmatrix} s_{11} & \cdots & s_{1m} \\ \cdot & & \cdot \\ \cdot & & \cdot \\ \cdot & & \cdot \\ s_{n1} & \cdots & s_{nm} \end{pmatrix} = \begin{pmatrix} e_1 & \ldots & e_m \\ s_{21} & & s_{2m} \\ \cdot & & \cdot \\ \cdot & & \cdot \\ \cdot & & \cdot \\ s_{n1} & \ldots & s_{nm} \end{pmatrix},$$

where the first row of s indicates the direct plus indirect requirements of the primary energy source. Then, for a given vector of final demand, y, the total primary input requirements are given by vector w:

$$w = \begin{pmatrix} y_1 e_1 + y_2 e_2 + \ldots + y_m e_m \\ y_1 s_{21} + y_2 s_{22} + \ldots + y_m s_{2m} \\ \cdot \\ \cdot \\ \cdot \\ y_1 s_{n1} + y_2 s_{n2} + \ldots + y_m s_{nm} \end{pmatrix}.$$

We can use the expressions

$$\begin{array}{lcl} r_{11} = y_1 e_1 & & r_1 = y_2 e_2 + \ldots + y_m e_m \\ r_{12} = y_1 s_{21} & & r_2 = y_2 s_{22} + \ldots + y_m s_{2m} \\ \cdot & \text{and} & \cdot \\ \cdot & & \cdot \\ \cdot & & \cdot \\ r_{1n} = y_1 s_{n1} & & r_n = y_2 s_{n2} + \ldots + y_m s_{nm} \end{array}$$

to define the symbols used in eq. (1). Note, moreover, that the same calculation can be reperformed using several alternative techniques for each of the sectors—i.e., alternative column values for a and q. This can yield a continuous production function F in the limit.

[8] Here it is assumed that energy is produced in a process that is completely separate from other outputs so that r_{i1} can be measured unambiguously because there are no common input costs.

used by it.[9] Here and throughout the remainder of the paper, we will take p_1 to be a constant; that is, we will assume that the United States is a price taker on the world energy market. This permits us to avoid the complications of analysis of strategic price behavior that the United States might want to consider. However, a bit will be said on this subject later.

Assume that input prices either are given exogenously or are for some other reason beyond the control of the firms that use them. Moreover, whether or not input-using firms are competitive, assume that all, or at least a substantial number, of them are cost minimizers, that is, that they seek to produce whatever output vector they select at minimum total cost. Then, as is well known, they will purchase those input quantities that satisfy

$$p_j/p_1 = F_{1j}/F_{11} = \partial r_{11}/\partial r_{1j} = \partial r_1/\partial r_j, \tag{3}$$

where $F_{1j} \equiv \partial F/\partial r_{1j}$, and so forth. Consequently, the profitability criterion, (2), can be rewritten as

$$\begin{aligned} y_1 \geqslant r_{11} + \sum_{j=2}^{n} r_{1j}F_{1j}/F_{11} &= r_{11} + \sum_{j=2}^{n} r_{1j}\partial r_{11}/\partial r_{1j} \\ &= r_{11} + \sum_{j=2}^{n} r_{1j}\partial r_1/\partial r_j. \end{aligned} \tag{4}$$

III. Tendency of Total Profit to Underestimate Total Net Energy Yield

The equivalence of (2) and (4) immediately suggests the relation between unprofitability and net energy loss, for

$$r_{1j}\partial r_{11}/\partial r_{1j} = r_{1j}\partial r_1/\partial r_j \tag{5}$$

appears to be an indicator of the energy opportunity cost of input j used in the energy manufacturing process: It is the quantity of input j used multiplied by the marginal quantity of energy input for which a unit of j can be substituted, ceteris paribus; that is, it is r_{1j} multiplied by the marginal rate of substitution of j for energy. Unfortunately, (5)

[9] One should note that here a crucial feature of energy production enters, namely, the fact that both the output and one of the inputs are really the same commodity. This is what permits us to assign the same price, p_1, to both the energy input, r_{11}, and the energy output, y_1. This is rather reminiscent of Quesnay's *Tableau économique*, where the output of agriculture was precisely the same product as its input—corn. The fact that energy is both an input and the output in this process permits us to divide (2) through by the product price, p_1, to obtain a ratio of input prices, p_j/p_1, which can be interpreted with the aid of efficiency condition (3). For noneconomist readers we may note that the standard relation (3) is obtained directly by minimization of total cost, $\Sigma p_j r_{1j}$, subject to the production function constraint $F(\cdot) = 0$.

is not the opportunity cost we are seeking, and we will show now that the correct figure is likely to be lower than (5) in many plausible circumstances.

The figure we really want corresponds to the saving in energy which would be permitted not by a small reduction in the use of j in activity 1 but by total elimination of the use of j for this purpose. To permit us to describe the issue with the aid of a two-dimensional diagram first, assume there are only two inputs, energy, r_1, and another input, r_2. The upper part of figure 1 represents the manufacture of all goods other than energy, and yy' is the isoproduct locus corresponding to the initial equilibrium, with point E representing the initial input combination.

Immediately below point E we have drawn the r_{11} axis of a graph representing the manufacture of energy, with E_1 its initial input vector, so that r_{12} is the initial quantity of input 2 used in the process. If all firms are cost minimizers, dr_1/dr_2, the slope of yy' at E, will equal dr_{11}/dr_{12}, the slope of y_1y_1' at E_1.

Transfer of all of r_{12} from production of energy to production of other goods, holding the output of the latter constant, will move the economy from point E to E' on curve yy' in the upper figure, thereby reducing the use of input 1 by $SE' = \Delta r_1 = r_{12}\Delta r_1/\Delta r_2$. This is the true energy opportunity cost of the use of input 2 in production of energy.

In general terms, to determine the energy opportunity cost of input j, we must take the quantity of this input used in energy production to fall from r_{1j} all the way to zero. Simultaneously, we must take that amount of input j, $\Delta r_j = r_{1j}$, to be transferred into activities elsewhere in the economy and determine the quantity of energy that Δr_j can replace in those other industries, holding all those outputs constant. The opportunity-cost figure we are seeking, then, is $r_{1j}\Delta r_1/\Delta r_j$, the decrement in use of energy permitted by the incremental use of $\Delta r_j = r_{12}$ elsewhere in the economy. Thus our net energy criterion becomes, instead of (4),

$$y_1 \geqslant r_{11} + \sum_{j=2}^{n} r_{1j}\Delta r_1/\Delta r_j. \tag{6}$$

This relationship constitutes our definition of the net energy yield of an activity: It measures whether the energy output of the activity is greater than the direct and indirect use of energy inputs plus the sum of the energy opportunity costs of all other inputs used in energy production if all other activities in the economy remain constant.

Let us now see the relation of net energy criterion (6) to the profitability criterion (4) and, in particular, to the terms $r_{1j}\partial r_1/\partial r_j$ which it employs. In figure 1 draw the tangent ET to yy' at E. Then, clearly,

$$r_{12}\partial r_1/\partial r_2 = ST > SE' = r_{12}\Delta r_1/\Delta r_2. \tag{7}$$

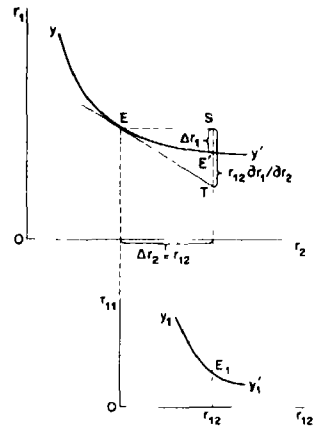


FIG. 1.—Calculation of an input's potential energy opportunity cost

This inequality will hold whenever the production surface is quasi-concave so that the isoproduct curve will be convex to the origin, as is yy'. For then, with the slope of yy' diminishing in absolute value as one moves to the right, $\Delta r_1/\Delta r_2$ must always be smaller, absolutely, than dr_1/dr_2. Intuitively, this is so because as more and more r_2 is used to replace energy in general production, one can be expected to be left with energy uses which become ever increasingly difficult to replace, for efficiency in the substitution process requires the easy energy-replacement opportunities to be used up first.

Thus, where (as usually assumed) the isoproduct curve has the shape of yy' the relationship equivalent to (7) can be expected to hold for all inputs, j. It then follows that the RHS (right-hand side) of (6) can generally be expected to be smaller than that of (4). Since (4) is the test of profitability, while (6) is the test of net energy yield, it follows that the latter may frequently be easier to pass than the former. That is, a new energy manufacturing process may be unprofitable and yet, nevertheless, yield a positive net energy output. However, the difference can be expected to be small. The reason is that what is at issue is the difference between marginal and average opportunity cost, not in the energy-producing industry but in the remainder of the economy taken as a whole. Since the quantities of input involved are likely to be small relative to the economy's total use of these inputs, the range of the isoproduct hypersurface traversed can be expected to be negligible. Thus the change in slope and, consequently, the difference between (4) and (6) should normally be small.[10]

[10] It must be reemphasized that the calculation is all predicated on the assumption that all outputs other than domestic energy production remain constant. It also ignores any price effects that can be expected to follow from a large change in an energy-producing activity.

We may then arrive at:

Proposition 1: While unprofitability does constitute grounds for suspicion that a new energy process will be a net energy loser, it cannot be taken necessarily to reject that process on net-energy grounds.

However, even this guarded conclusion suggests a judgment more sanguine than the situation warrants, as will be shown in the next section.

IV. Marginal Unprofitability Implies Negative Marginal Yield of Net Energy

We will show now:

Proposition 2: If firms in the economy minimize the cost of producing their output vectors, a marginal increase in energy output from any process that has a negative profit yield must also have a negative potential marginal yield of net energy when other outputs are held constant.

This conclusion has policy implications which are extremely disturbing:

Corollary: If a firm that invests in an energy-producing process can capture all of its profit yield (i.e., if there are no externalities or similar market failures), any expansion in the manufacture of energy (beyond the quantity that maximizes profits) that is induced by subsidy unaccompanied by a specially designed program of taxes on energy use elsewhere in the economy must be excessive in the sense that (at least) a small *reduction* in that energy output can *increase* the economy's energy stocks if all other activities in the economy are unaffected.[11]

Put alternatively, since such an expansion in the energy-producing process may well lead indirectly to a reduction in other activities in the economy, some energy may well be saved as a consequence. But any saving in energy would have been greater if those other activities had been reduced in the same way without the expansion in the energy-production process that is subsidized.

In other words, whenever a subsidy to an energy-producing activity leads to its expansion beyond the level to which it would be brought by market forces, it can be expected to have a negative (marginal) net energy yield in this sense. That is, under the circumstances postulated any subsidy may make matters worse, not just in cost-benefit terms, as we might normally expect such a subsidy to do, but in terms of physical energy supplies.

[11] This assumed implicitly that the profit surface is concave (hill shaped) and that in the absence of subsidy the firm will operate at the point of maximum profit or beyond it (as it will, e.g., under sales maximization). Then the subsidy-free output must correspond to a downward-sloping portion of the relevant cross section of the profit hypersurface, and the result applies.

The proof of this disturbing proposition is straightforward. Negative marginal profitability (on our parametric price assumption) is described by[12]

$$p_1 dy_1 < \Sigma p_j dr_{1j}, \tag{8}$$

but by (3), (8) can be rewritten as

$$dy_1 < dr_{11} + \sum_{j=2}^{n} dr_{1j} \partial r_1 / \partial r_j, \tag{9}$$

which tells us that the marginal net energy yield of output y_1 is negative. Q.E.D.

One remarkable feature of this proof is that it does not rest on a restrictive premise such as the presence of perfect competition. The key assumption which underlies (3) is that firms minimize the total cost of whatever output they produce. But this holds for a wide variety of market forms—for example, profit-maximizing oligopolists, monopolists or monopolistic competitors, and firms which maximize total revenue (sales)[13] or growth rather than profits.[14]

Nor is it necessary that this premise be satisfied throughout the economy, as long as it holds for the sectors where the inputs will go if they are not used in energy production, and, in particular, (3) does not have to be satisfied in the energy-producing sector itself. That is, we do not have to assume that the economy is initially operating efficiently, for the opportunity cost, as we have seen, is calculated not in the energy-producing sector but in the remainder of the economy where input j can be substituted for energy use. Thus, if somewhere else in the economy least-cost production occurs and satisfies $p_j/p_1 = \partial r_1/\partial r_j$, a trade-off on the terms indicated is indeed possible.[15] And if

[12] This assumes that energy price, p_1, is determined on the world market and is therefore uninfluenced by the output of the domestic industry that manufactures energy. If this is not so, $R_a = p_1(dy_1 - dr_{11})$ should be replaced by $R_b = d(p_1 y_1 - p_1 r_{11}) = (y_1 - r_{11})dp_1 + p_1(dy_1 - dr_{11})$. With a negatively sloping demand curve, when $dy_1 > 0$ then $R_b < R_a$. This means that when marginal profit is zero, the marginal net energy yield will be positive so that some subsidy to the output of a profit-maximizing producer of energy will in this case contribute to the economy's energy stocks. However, since it seems unlikely that the magnitude of the output of any one producer of synthetic fuels, e.g., will be sufficient to have a significant effect on international oil prices, this point does not seem particularly important for policy.

[13] However, our result does have special implications for firms with objectives such as sales maximization which drive them beyond the point of maximum profits even in the absence of subsidy. If energy production is in the hands of such a firm, our result shows that even without subsidy, its energy output will be carried to a point where its marginal net energy yield is negative, so that a restriction in its output will actually increase the economy's energy stocks. Thus a tax on such an activity, rather than just a zero subsidy, is called for, at least in theory.

[14] Thus this differs from the usual conclusion that subsidy will reduce welfare in that our current conclusion is expressed in physical rather than psychological (utility) terms and, unlike the usual result, does not require universal perfect competition.

[15] Thus the neoclassical assumption of substitutability at the margin, as implied in

this relationship holds in those sectors where the resources will actually be used, such a trade-off will in fact occur.

V. On the Reality of the Energy Opportunity Costs

We must be careful not to exaggerate the strength of our result. An energy opportunity cost clearly does represent a *potential* energy loss, and in that sense an expansion of an energy-production process *does* represent an energy loss relative to what would have been achieved if its resources had instead been used in an optimal program to increase energy stocks. But in reality, if the energy-producing activity were reduced from y_{1s}, its magnitude under a subsidy, to $y_{1\pi}$, its profit-maximizing level, the inputs released as a result might very well not be used in a way that maximizes their contribution to the energy program. First, they might not be used efficiently and, second, they would probably be used in part for other purposes. They *would* be used in an efficient manner if (3) were to hold everywhere at either level of y_1, that is, if all firms in the economy were to produce whatever output vectors they choose to supply at the minimum cost for that vector. But if that is not so, the actual energy loss resulting from the increased use of other inputs needed to raise y_1 to y_{1s} may well be less than our calculations indicate. In addition, whether or not they were reallocated efficiently, in practice, only some portion of those reallocated resources would be used as a substitute for oil. That is why the corollary to proposition 2 has been put rather guardedly.

Nevertheless, those losses will almost certainly be real (and very possibly substantial, as is suggested by the material reported in n. 6 above). That is, if the energy-producing activity is in fact reduced and the corresponding input quantities are thrown on the market, the forces of supply and demand can be expected to cause a readjustment in relative input prices which induce the employment of at least some of those inputs elsewhere in the economy. Those other inputs will then in fact serve (in part) as energy substitutes. Thus the marginal energy opportunity cost of the labor, materials, and other primary inputs used in energy production will represent some *actual* loss in energy stocks and not a remote possibility of substitution which is perhaps unlikely to occur.[16] Interpreted in this sense, our propositions no

---

$p_j/p_1 = \partial r_1/\partial r_j$, need not hold throughout the economy. For the weaker interpretation of our result (see next section) to be valid, all that is required is that *somewhere* in the economy the resources dr_1, absorbed by the energy sector when its output is increased by dy_1, be substitutable for energy. It seems hard to imagine that this weak premise will ever fail to be satisfied.

[16] See n. 2 above. We may inquire briefly about the implications of violations of some of the premises underlying proposition 2 In particular, we will consider (*a*) the effects of the presence of specialized inputs in the energy-using industry which are not usable elsewhere, (*b*) monopsony power in the input-using industry, and (*c*) cases where

longer *guarantee* that a subsidy will always cause an incremental net energy loss, but they nevertheless leave it under a dark shadow of suspicion.

We can put the matter another way by recognizing explicitly that there are two different opportunity-cost concepts, which we may refer to as the *potential* and the *actual* opportunity losses. In our case, the former is the maximal amount of petroleum that can be saved by substitution of some quantity of another resource, Δr_j, while the latter refers to the amount of oil that will actually be saved by the use to which Δr_j will in fact be put when it is released from a subsidized activity. Our formal propositions evaluate the opportunity cost in terms of potential opportunity loss, but there is good reason for policymakers to be concerned with the actual opportunity loss, whose magnitude, while undoubtedly substantial, is certainly lower than the potential-loss figure.

We can now use a diagram to suggest another way of looking at the analysis, which may help to remove any air of sleight of hand that may characterize our results.[17] In figure 2, two production-possibility frontiers are represented for the economy, showing the variables, net energy outputs and other outputs, on the axes. Here YY^0 represents the production-possibility frontier relevant without the subsidy for the energy supply technique, while $Y'Y''$ represents the production-possibility frontier relevant with the energy supply subsidy. This latter frontier is drawn below and to the left of the former since the subsidy reduces economic efficiency. Here the reduction in economic efficiency can be approximated by the following integral:

$$\int_{E_0}^{E_1} S(E)dE,$$

where $S(E)$ represents the subsidy offered per unit of supply, E represents the quantity of energy produced using the technique in

management derives utility from the hiring of a quantity of input j greater than that which minimizes cost. The first of these works to undermine proposition 2, since such specialized inputs enter the incremental profit expression as a money cost though they incur no energy opportunity cost. Thus they may cause failure of the profit test even when the net energy test is passed However, the next type of violation of our premises can be expected to work the other way, strengthening proposition 2. Monopsony power over input j means that in equilibrium $\partial r_1/\partial r_j = mc_j/p_1 > p_j/p_1$, since we expect the supply price of j to be rising (here mc_j is the marginal cost of purchases of j to the firm that uses it as an input). Hence the true opportunity cost of j, $\partial r_1/\partial r_j$, will be greater than p_j/p_1, the corresponding figure in the profit equation. Finally, excessive purchases of j, because they provide utility to management, need not affect our result. With a given stock of resource j it is impossible for this to be a universal phenomenon. But if it occurs only in a limited sector of the economy, elsewhere j will become scarcer, its relative marginal yield will rise, but its price will rise correspondingly, and $p_j/p_1 = \partial r_1/\partial r_j$ will continue to be satisfied by firms in the remainder of the economy (i.e., those which do not buy excessive amounts of j).

[17] We are grateful to James Sweeney and Robert Willig for suggesting this way of looking at the matter. The next several paragraphs follow a draft by Sweeney.

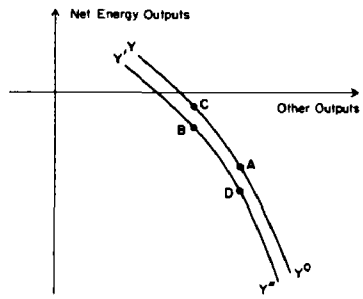


FIG. 2.—Production-possibility frontiers with and without energy supply subsidy

question, E_0 represents its initial output, and E_1 represents the output under the subsidy. In figure 2, A represents the initial output combination in the absence of any subsidy, and B represents the output combination with the subsidy. These points have been drawn under the assumption that the subsidy in fact increases net energy outputs and decreases the outputs of other goods and services. Point C represents a combination of net energy outputs and other outputs available if a perfectly efficient mechanism, such as a tax on the imports of all energy, had been instituted. The difference in net energy outputs between C and B is what we call the potential net energy loss associated with this subsidy. Note that, although the subsidy program can increase the actual net energy output, there is a loss in potential net energy output relative to what could be obtained with a perfectly efficient program.

Alternatively, the distance from A to D can be viewed as the potential net energy output loss associated with a subsidy program. This is the net energy loss if the economy is managed in such a way that all other outputs remain constant.

Thus we see that even though a subsidy program may increase actual net energy output, it can be expected to lead to a decrease in potential net energy output.

VI. Promise of Future Profitability: A Dangerous Fallacy

Before we turn to our overall conclusion, it is important to deal with a critical error in reasoning which the preceding analysis may well elicit. Undoubtedly it will suggest the response that many new energy sources, while admittedly unprofitable today (both at the margin and overall), should become viable in the foreseeable future if enough R & D is devoted to the task. This is probably true, but it weakens neither the analysis nor its policy implications, because the issue is not whether they will ultimately begin to yield positive profits, but whether those profits can be expected to be sufficient to compensate

for the initial losses. To drive home the point, consider each of the two possibilities explicitly.

Case a: If future profits are insufficient to make up for earlier losses at the margin, our analysis still shows that over its lifetime the investment will be a potential net energy loser. It will yield some net energy gains in later years, but not enough to make up for the resource depletion it caused earlier. To see how this is implied by our previous analysis, we need merely interpret the quantities in our equations to represent discounted present values of the corresponding future streams, and the result follows at once: The present value of the stream of future profits serves as the required indicator of the present value of the stream of values of the future net energy yields, and propositions 1 and 2 continue to apply.[18]

Case b: Where expected future profits do make up for earlier losses, the corresponding net-energy tests will, of course, be passed. But in the absence of capital market imperfections and externality problems, there should then be no need for government subsidies. Private enterprises should, in these circumstances, be expected to invest adequately in the project. After all, any firm that invests in an innovation normally expects to lose money initially and to make it up later. There is simply nothing unusual in such cases.

Thus we conclude that the likelihood that with sufficient research outlays a new energy source can be made viable financially at some point in the future does not weaken the presumption that subsidies to such projects will lead to a net loss in the economy's potential net energy stocks.

Here it should be noted, incidentally, that the research process itself has its energy costs—direct, indirect, and those that take the form of energy opportunity costs. (One fusion-research project, e.g., is reported to use as much electricity as a medium-sized American city.) It may also be appropriate at this point to take note of another potentially misleading argument on our subject. It is tempting to assert that, because the energy issue is not exclusively a matter of growing scarcity of a resource but involves freedom of the nation's foreign policy from the threat of international blackmail, it may well pay society to invest in energy-production projects to an extent that is unprofitable for the private sector. By itself, this assertion is legitimate enough, but unfortunately there is no way such investments will help to deal with the international blackmail problem if the projects undertaken are net energy losers. In short, we are by no means advocating adherence to a profitability criterion for its own sake. We are merely suggesting that

[18] It can, however, be argued (and, indeed, it has been) that it is inappropriate to discount a flow of benefits to future generations. This is not the place to go into the complexities of this difficult issue—the proper standards of justice in intergenerational distribution.

profit can often serve as a useful indicator of the net energy contribution promised by such an undertaking. If its net energy yield over its lifetime will be negative, there is no way it can be justified on geopolitical grounds any more than it can be rationalized in economic terms.

VII. Heterogeneity of Energy Sources

There is another assumption underlying our analysis which can easily lead to a second erroneous conclusion on policy. We have assumed all forms of energy to be homogeneous, but in fact they are not. Coal is very different from oil in terms of the uses to which it can easily be put, in transportability, in terms of environmental effects, and so forth. Such differences may sometimes be able to justify the adoption of projects which are (marginal) net energy losers. Suppose one type of energy (petroleum) is in short supply while another (coal) is very abundant. An effective process to transform coal into oil may then be worthwhile even though the nation's inventory of British thermal units is reduced in the process.

Heterogeneity may take yet another form—that of differences in the dates at which different energy supplies become available. Suppose we were to believe that within a century, fusion will supply all our energy needs safely, abundantly, and cheaply. The energy supplies which will not be used before the twenty-second century may become valueless or nearly so. If this is, for example, true of much of our coal reserves, a net energy loss involved in the trade of coal tomorrow for liquid fuel today may well be justifiable.

But the conclusion is emphatically not that the net energy considerations emphasized in this paper can be safely ignored. Indeed, it is remarkable that the robustness of our basic theorem (proposition 2) extends it to cover the issue of heterogeneity of energy sources. That is, we will see now that in the absence of externalities, a subsidy to a particular form of energy (such as current output of petroleum from coal) must actually reduce the potential net output of that form of energy (it must reduce net petroleum stocks even if other forms of energy are used as an input in the process). In showing this result, we can easily deal with an economy having any arbitrary number of forms of energy. But the logic of the argument as well as the nature of the formal analysis can be brought out most clearly by assuming that there are only two forms of energy, one of which (call it "petroleum") is directly substitutable for imported energy, while the other (call it "coal") is not directly substitutable for imported fuel and is available in large amounts which may never be used up.[19] We then have:

[19] Obviously, proposition 3 will not hold if coal is a homogeneous and totally free economic good which can be substituted for petroleum without cost. But coal does cost

Proposition 3: Under the conditions of proposition 2, the expansion of any petroleum-producing activity which involves a negative marginal profit yield will involve a net reduction in petroleum stocks at the margin, even if substantial amounts of coal are employed in the process, if output of coal and other outputs are held constant.

That is, such an unprofitable increase in petroleum output may well exacerbate the economy's energy importation problem, involving increased use of coal as a substitute for petroleum!

At first glance this result may be even more implausible than proposition 2 itself, but its proof is simply a matter of trivial reinterpretation of the steps of the proof of proposition 2. For this purpose we need merely consider y_1 to be manufactured petroleum (e.g., oil extracted from shale or from coal), r_1 and r_{11} to be the inputs of petroleum alone (rather than total energy input), and r_2 and r_{12} to be coal inputs. Then (10) immediately constitutes our result, for it tells us about the net marginal yield of commodity 1 (petroleum) alone, regardless of the quantities of input 2 (coal) or other resources used to produce it.

One may well argue vis-à-vis this proposition that it is unreasonable to assume that a subsidy, say, to conversion of coal to oil will leave other outputs unchanged, for it will surely affect at least the gross output of coal and perhaps the gross outputs of other energy sources. Nevertheless, our interpretation of proposition 2 in relation to this issue remains valid here. That is, the petroleum saved by a reduction in other activities in the economy may offset the oil lost by the expansion of the subsidized activity, but in that case the gain in oil inventories would have been still greater if those other outputs had contracted in the same way without the subsidy-induced increase in conversion of coal into oil.

However, in this case the behavior of substitute energy sources is actually apt to exacerbate the problem rather than relieve it: (*a*) The output of other forms of energy used as inputs in the subsidized process (the coal in our example) will expand and therefore increase their own use of oil as an input, thus tending to reduce the overall petroleum yield;[20] (*b*) the outputs of other forms of energy (say, electricity) may be reduced because subsidized oil becomes somewhat cheaper, but if this is so it will only occur because oil consumption is increased and substituted for other types of energy. On both scores,

money to extract, and it varies in quality and ease of extraction. Hence, it has a nonzero price per Btu which can be expected to rise as coal stocks are used up.

[20] It is, of course, true that in this case the available quantity of coal may be increased, but there are two reasons why it will not help to solve the problem: First, by hypothesis "coal" is used here to represent an input which, unlike oil, is not in short supply. Otherwise, there is nothing gained by sacrificing 10 Btu of coal for 5 Btu of oil, the issue we are now discussing. Second, by proposition 2, the subsidy must still have a negative potential net energy yield of coal and oil together as well as in oil alone (proposition 3).

then, the net energy loss in the subsidized production of oil is likely to be aggravated by the response in output of other types of energy.

VIII. Oil Tariff as Stimulus to Supply

We began our discussion by contrasting the net energy consequences of a targeted subsidy program with those of a tariff on imported oil. It is time to turn to the latter. Obviously, a tariff will tend to reduce oil use by both consumers and industry as well as stimulate energy supplies from both conventional and innovative sources.

Standard economic analysis also suggests that such a tariff will accomplish all this in an efficient manner.[21] If the magnitude of the tariff corresponds to the size of the marginal detrimental externality represented by increased dependence on OPEC, in a competitive world it will move the economy to a Pareto optimal point on the efficiency frontier. This point will involve larger oil stocks and smaller outputs of at least some other commodities than would otherwise have been produced. All this would, of course, be accompanied by some reduction in the consumption of those other goods by households and other consumers.

Nevertheless, our analysis may well appear to have shown that this benefit of the tariff is likely to be offset in part by its supply consequences. To the extent that it serves, in effect, as a subsidy to domestic production, it seems likely to lead to changes in output whose net energy yields are negative, with the tariff protecting producers from the reductions in profit that would otherwise follow. This view of the matter is, however, incorrect. Since a tariff raises the cost of imported energy everywhere, it will not only increase the gross output of domestic production of energy in export-substitutable forms; it will also cause substitution of other inputs for use of such types of energy elsewhere in the economy. This will cause adjustments in the energy opportunity costs of those inputs, which will mean that any profitable expansion in domestic energy production resulting from the tariff will in fact have a positive net energy yield. This is because the increased substitution, throughout the economy, of other inputs for oil will cause the marginal rate of substitution (MRS) to fall; that is, an additional unit of such an input will be able to replace less oil than it could before. In this way the MRS between oil and any other input elsewhere in the economy will remain equal to its marginal product in innovative petroleum-producing activities.

In terms of figure 2 the point is that the tariff will move the economy from a point such as A to one such as C, with an increase in oil output and a loss in efficiency, while the subsidy will, instead, move us to a point such as B, causing the potential net energy loss BC.

[21] For an excellent discussion along these lines, see Schmalensee (1979).

Somewhat more formally, we see that the pretariff output of an innovative energy source will be smaller than the amount that maximizes profits after the imposition of the tariff. As we have seen, an increase in output that offers a positive marginal profit must also have positive marginal net yield of energy. It follows that the net energy yield of the expansion in output of manufactured energy after imposition of the tariff must be positive.

Nevertheless, there are some reservations which must be expressed about the tariff approach. First, if a tariff induces an increase in consumption from domestic sources, it may simply hasten the process of depletion of our oil inventories and speed the economy's growth in vulnerability to the cartel!

One can also think of a case in which subsidy of a net energy losing project with high sunk costs (e.g., high research costs but low variable costs) may conceivably be superior to a tariff. If, once the resources are expended, it constitutes a credible threat which prevents the cartel from raising its price in the future, it may be worth the outlay. It may be superior to a tariff which will also lead to domestic investment outlays on energy production and, in addition, result in higher prices to consumers.[22]

IX. Concluding Comment: Subsidies Which Do Increase Energy Supplies

We began by denying that we were opposed to all programs involving subsidies to new energy sources. Our objective, rather, is to make absolutely clear how easily an ill-considered subsidy program can backfire and may actually serve to deplete the economy's energy supplies. But we may now seem to have proved too much—to have shown that *any* subsidy can be presumed to cause energy losses. In fact, we have not done so. Rather, properly interpreted, our results indicate the conditions under which such outlays can contribute energy resources. This is made possible by the role of externalities that may produce cases in which a project's net energy yield offers no corresponding net profits to the firm that undertakes it.

There is no point in attempting to provide an exhaustive list of the ways in which such sources of market failure may arise. We will only mention two which, we suspect, are of prime importance here and which should illustrate the sort of projects for which subsidies can be justified. The two externalities which will be discussed are those resulting from differences between private and social risk and those which arise from the free-rider problems engendered by investment in research.

[22] We are indebted to Robert Willig for pointing out this possibility to us.

There are at least two reasons for which the social risks of an investment may be smaller than those incurred by a private firm which undertakes it, so that the investment may be desirable for society even though it is unattractive to private enterprises. The first is the well-known argument of Arrow and others which points out that since at any given time the economy contains a substantial number of investments which incur risks that are to a considerable degree independent, the "law of large numbers" reduces the risk to society of an additional project below what it is to a private entity which has a considerably smaller number of projects under its control.

Second, financial risk is likely to be more threatening to the individual owner of an enterprise than it is to society. If the project goes bankrupt but continues to operate in receivership, the investment is by no means lost to the community, though it is lost completely to the initial investor. So, on these grounds, too, new energy projects, which are indeed risky, may incur social costs and, hence, energy costs lower than their energy yield even if they are unattractive to private investors.

In practice, perhaps the most important justification for subsidy to particular energy production projects is the free-rider problem which accompanies all applied research. Research almost always generates information which is useful to others besides those who have conducted or financed it, and there is often no way to make those others bear an appropriate share of the cost. Thus, taking all beneficiaries together, the total yield of the research may well exceed its costs, and our analysis can be shown to suggest that an energy project which does so can be presumed to have a positive net energy yield.[23] Nevertheless, because some of the beneficiaries do not make a corresponding financial contribution toward the cost of the research, such a project may well be unattractive to any private entrepreneur.

The patent system is, of course, intended to deal with this free-rider problem, and to some extent it does. However, some research results, particularly those that qualify as products of "basic research," are not patentable. Others are early steps in a development process which will occupy many years—so many that the period of patent protection is exceeded. And even results which are eligible for patent protection may nevertheless throw off substantial indirect benefits to free riders who will be helped by them to produce substitutes for the patented item. This is frequently suggested as the explanation for the fact that

[23] This is fairly obvious if the free riders use their unpaid-for knowledge to produce more energy; their energy output must be added to that of those who conducted the research. However, the argument applies even if the free riders use the knowledge in other economic activities, since the product of those activities will also have an energy opportunity value. For example, if the new knowledge reduces the inputs needed to produce a unit of some other product, the inputs saved can be used as a substitute for energy elsewhere in the economy, in the manner already discussed.

many firms choose not to patent the results of their R & D activities, thereby avoiding the disclosure required for a patent application.

These are probably the prime reasons no private firms will undertake the research outlays required for a protracted and risky development process such as that required for very innovative energy sources like nuclear fusion, which is generally not expected to yield a financially viable product until we are well into the twenty-first century.

In conclusion, let us emphasize again that even if a benign energy source such as solar energy does not pass the test that qualifies it for subsidies under our criterion, it may nevertheless be socially desirable to encourage its production. But a tariff on oil importation or a tax on the consumption of oil would appear to be the appropriate way to do so.

References

Baron, Seymour. "Energy Cycles: Their Cost Interrelationship for Power Generation." *Mechanical Engineering* 98, no. 6 (June 1976): 22–30.

———. "Solar Energy—Will It Conserve Our Non-Renewable Resources?" Mimeographed. Oradell, N.J.: Burns & Roe, 1978.

Berry, R. Stephen. "Energy Analysis for Energy Policy." *J. Bus. Admin.* (Special Issue: Energy Policy: The Global Challenge) 10, nos. 1 and 2 (Fall 1978–Spring 1979): 83–105.

Center for Advanced Computation, Energy Research Group. "Energy Flow through the United States Economy." Wall chart, Univ. Illinois at Urbana-Champaign, 1975.

Fels, Margaret F., and Williams, G. Pierce. "Total Energy Requirements to Construct and Operate Heavy Rail Transit Systems." Transportation Report no. 78-TR-25, Princeton Univ., Transportation Program, April 1978.

Hunt, Lee P. "Total Energy Use in the Production of Silicon Solar Cells from Raw Materials to Finished Product." In *12th IEEE Photovoltaic Specialists Conference—1976*. New York: Inst. Electrical and Electronics Engineers, 1976.

New York Times. "A Way to Save Energy" (February 24, 1980).

Perry, Alfred M., et al. "Net Energy Analysis of Five Energy Systems." ORAU/IEA (r)-77-12. Oakridge, Tenn.: Inst. Energy Analysis, Oakridge Assoc. Univ., September 1977.

Ritz, Philip M. "The Input-Output Structure of the U.S. Economy, 1972." *Survey Current Bus.* 59 (February 1979): 33–43, 72.

Ritz, Philip M.; Roberts, Eugene P.; and Young, Paula C. "Dollar Value Tables for the 1972 Input-Output Study." *Survey Current Bus.* 59 (April 1979): 51–72.

Schmalensee, Richard. "Appropriate Government Policy toward Commercialization of New Energy Supply Technologies." Working Paper no. MIT-EL 79-052WP, Massachusetts Inst. Tech., Energy Laboratory, October 1979.

Shabecoff, Philip. "Experts Foresee Economic Shift with Less Energy and More Jobs." *New York Times* (February 18, 1980).

A Rational Theory of the Size of Government

Allan H. Meltzer and Scott F. Richard

Carnegie-Mellon University

In a general equilibrium model of a labor economy, the size of government, measured by the share of income redistributed, is determined by majority rule. Voters rationally anticipate the disincentive effects of taxation on the labor-leisure choices of their fellow citizens and take the effect into account when voting. The share of earned income redistributed depends on the voting rule and on the distribution of productivity in the economy. Under majority rule, the equilibrium tax share balances the budget and pays for the voters' choices. The principal reasons for increased size of government implied by the model are extensions of the franchise that change the position of the decisive voter in the income distribution and changes in relative productivity. An increase in mean income relative to the income of the decisive voter increases the size of government.

I. Introduction

The share of income allocated by government differs from country to country, but the share has increased in all countries of western Europe and North America during the past 25 years (Nutter 1978). In the United States, in Britain, and perhaps elsewhere, the rise in tax payments relative to income has persisted for more than a century (Peacock and Wiseman 1961; Meltzer and Richard 1978). There is, as

We are indebted to Karl Brunner, Dennis Epple, Peter Ordeshook, and Tom Romer for many helpful discussions and to the participants in the Carnegie-Mellon Public Economics Workshop, an anonymous referee, the editor, and the Interlaken Seminar for constructive comments on an earlier version.

[*Journal of Political Economy*, 1981, vol. 89, no. 5]
© 1981 by The University of Chicago. 0022-3808/81/8905-0008$01.50

yet, no generally accepted explanation of the increase and no single accepted measure of the size of government.

In this paper, the budget is balanced.[1] We use the share of income redistributed by government, in cash and in services, as our measure of the relative size of government and develop a theory in which the government's share is set by the rational choices of utility-maximizing individuals who are fully informed about the state of the economy and the consequences of taxation and income redistribution.[2]

The issues we address have a long intellectual history. Wicksell (1958) joined the theory of taxation to the theory of individual choice. His conclusion, that individual maximization requires government spending and taxes to be set by unanimous consent, reflects the absence of a mechanism for grouping individual choices to reach a collective decision. Following Downs (1957), economists turned their attention to the determination of an equilibrium choice of public goods, redistribution, and other outcomes under voting rules that do not require unanimity.

Several recent surveys of the voluminous literature on the size or growth of government are now available (see Brunner 1978; Peacock 1979; Aranson and Ordeshook 1980; and Larkey, Stolp, and Winer 1980).[3] Many of the hypotheses advanced in this literature emphasize the incentives for bureaucrats, politicians, and interest groups to increase their incomes and power by increasing spending and the control of resources or rely on specific institutional details of the budget, taxing, and legislative processes. Although such studies contribute to an understanding of the processes by which particular programs are chosen, they often neglect general equilibrium aspects. Of particular importance is the frequent failure to close many of the models by balancing the budget in real terms and considering the effect on voters of the taxes that pay for spending and redistribution (see, e.g., Olson 1965; Niskanen 1971; and Hayek 1979). A recent empirical study by Cameron (1978) suggests that decisions about the size of the budget are not the result of "fiscal illusion," so the neglect of budget balance cannot be dismissed readily.

We differ from much of the recent literature in three main ways.

[1] All variables are real. There is no inflation. Budget balance means that redistribution uses real resources. Public goods are neglected.

[2] Ideally the size of government would be measured by the net burden imposed (or removed) by government programs.

[3] Larkey et al. (1980) include a survey of previous surveys. Recent surveys by Mueller (1976) and Sahota (1978) summarize recent contributions by Downs (1957), Musgrave (1959), Olson (1965), Niskanen (1971), Buchanan and Tullock (1972), Riker and Ordeshook (1973), and others to such related topics as the determination of equilibrium collective decisions and the effects of government policies on the distribution of income.

First, voters do not suffer from "fiscal illusion" and are not myopic. They know that the government must extract resources to pay for redistribution. Second, we concentrate on the demand for redistribution and neglect any "public goods" provided by government (see also Peltzman 1979). Third, we return to the earlier tradition of de Tocqueville ([1835] 1965) who associated the size of government, measured by taxes and spending, with two factors: the spread of the franchise and the distribution of wealth (property).[4]

Our hypothesis implies that the size of government depends on the relation of mean income to the income of the decisive voter. With universal suffrage and majority rule, the median voter is the decisive voter as shown by Roberts (1977) in an extension of the well-known work of Hotelling (1929) and Downs (1957). Studies of the distribution of income show that the distribution is skewed to the right, so the mean income lies above the median income. Any voting rule that concentrates votes below the mean provides an incentive for redistribution of income financed by (net) taxes on incomes that are (relatively) high. Extensions of the franchise to include more voters below mean income increase votes for redistribution and, thus, increase this measure of the size of government.

The problem with this version of the de Tocqueville hypothesis is that it explains too much. Nothing limits the amount of redistribution or prevents the decisive voter from equalizing incomes or, at a minimum, eliminating any difference between his disposable income and the disposable income of those who earn higher incomes. Incentives have been ignored. Higher taxes and redistribution reduce the incentive to work and thereby lower earned income. Once we take account of incentives, there is a limit to the size of government. To bring together the effect of incentives, the desire for redistribution, and the absence of fiscal illusion or myopia, we develop a general equilibrium model.

Section II sets out a static model. Individuals who differ in productivity, and therefore in earned income, choose their preferred combination of consumption and leisure. Not all individuals work, but those who do pay a portion of their income in taxes. The choice between labor and leisure, and the amount of earned income and taxes, depend on the tax rate and on the size of transfer payments.

The tax rate and the amount of income redistributed depend on the voting rule and the distribution of income. Section III shows how income redistribution, taxes, and the size of the government budget

[4] We are indebted to Larkey et al. (1980) for pointing out the similarity between de Tocqueville and the conclusion we reached in an earlier version and in Meltzer and Richard (1978). De Tocqueville's distribution of property finds an echo in the concerns about "mob rule" by the writers of the Constitution.

change with the voting rule and the distribution of productivity. A conclusion summarizes the findings and main implications.

II. The Economic Environment

The economy we consider has relatively standard features. There are a large number of individuals. Each treats prices, wages, and tax rates as givens, determined in the markets for goods and labor and by the political process, respectively. Differences in the choice of labor, leisure, and consumption and differences in wages arise solely because of differences in endowments which reflect differences in productivity. In this section, we extend this standard model to capture the salient features of the process by which individuals choose to work or subsist on welfare payments and show the conditions under which these choices are uniquely determined by the tax rate.

The utility function is assumed to be a strictly concave function, $u(c, l)$, for consumption, c, and leisure, l. Consumption and leisure are normal goods, and the marginal utility of consumption or leisure is infinite when the level of consumption or leisure is zero, respectively. There is no capital and no uncertainty.

The individual's endowment consists of ability to produce, or productivity, and a unit of time that he allocates to labor, n, or leisure, $l = 1 - n$. Individual incomes reflect the differences in individual productivity and the use of a common, constant-returns-to-scale technology to produce consumption goods. An individual with productivity x earns pretax income, y:

$$y(x) = xn(x). \tag{1}$$

Income is measured in units of consumption.

Tax revenues finance lump-sum redistribution of r units of consumption per capita. Individual productivity cannot be observed directly, so taxes are levied against earned income. The tax rate, t, is a constant fraction of earned income but a declining fraction of disposable income. The fraction of income paid in taxes net of transfers, however, rises with income.[5] There is no saving; consumption equals

[5] Reliance on a linear tax follows a well-established tradition. Romer (1975) analyzed problems of unimodality using a linear tax and predetermined government spending. Roberts (1977), using a linear tax and a predetermined budget, showed that the median voter dominates the solution if incomes are ordered by productivity. Linear tax functions are used also when the social welfare function is used to determine the optimal tax (see Sheshinski 1972). The degree to which actual taxes differ from linear taxes has generated a large literature. Pechman and Okner (1974) find that the tax rate is approximately constant. King (1980) writes that most redistribution in the United States and the United Kingdom comes from the transfer system, not from the tax system. Browning and Johnson (1979) show that conclusions about proportionality of the tax rate depend heavily on assumptions used to allocate the burden of indirect business taxes.

disposable income as shown in (2):

$$c(x) = (1 - t)nx + r, \qquad c \geq 0. \tag{2}$$

If there are individuals without any ability to produce, $x = 0$, their consumption is $r \geq 0$.

Each individual is a price taker in the labor market, takes t and r as givens, and chooses n to maximize utility. The maximization problem is:

$$\max_{n \in [0,1]} u(c, l) = \max_{n \in [0,1]} u[r + nx(1 - t), 1 - n]. \tag{3}$$

The first-order condition,

$$\begin{aligned} 0 = \frac{\partial u}{\partial n} &= u_c[r + nx(1 - t), 1 - n]x(1 - t) \\ &\quad - u_l[r + nx(1 - t), 1 - n], \end{aligned} \tag{4}$$

determines the optimal labor choice, $n[r, x(1 - t)]$, for those who choose to work. The choice depends only on the size of the welfare payment, r, and the after-tax wage, $x(1 - t)$.[6]

Some people subsist on welfare payments. From (4) we know that the productivity level at which $n = 0$ is the optimal choice is

$$x_0 = \frac{u_l(r, 1)}{u_c(r, 1)(1 - t)}. \tag{5}$$

Individuals with productivity below x_0 subsist on welfare payments and choose not to work; $n = 0$ for $x \leq x_0$.

Increases in redistribution increase consumption. For those who subsist on welfare, $c = r$, so $\partial c/\partial r = 1$. Those who work must consider not only the direct effect on consumption but also the effect of redistribution on their labor-leisure choice. The assumption that consumption is a normal good means that $\partial c/\partial r > 0$. Differentiating (4) and using the second-order condition, $D < 0$, in footnote 6 restricts u_{cl}:

$$\frac{\partial c}{\partial r} = \frac{u_{cl}x(1 - t) - u_{ll}}{-D} > 0. \tag{6}$$

Consumption increases with r for both workers and nonworkers provided consumption is a normal good.

The positive response of c to r takes one step toward establishing conditions under which we find a unique value of r that determines

[6] By assumption, u is strictly concave, so the second-order condition is negative and (4) defines a maximum. The second-order condition is $\partial^2 u/\partial n^2 = D = u_{cc}x^2(1 - t)^2 - 2u_{cl}x(1 - t) + u_{ll} < 0$.

the amount of earned income and amount of redistribution for each tax rate. The next step is to show that normality of consumption is sufficient to establish that earned income (income before taxes) increases with productivity.

Pretax income is

$$y(r, t, x) = xn[r, x(1 - t)]. \tag{7}$$

People who do not work, $x \leq x_0$, have $y = 0$ and $\partial y/\partial x = 0$. For all others,

$$\frac{\partial y}{\partial x} = n + x\frac{\partial n}{\partial x}. \tag{8}$$

The first-order condition (eq. [4]) yields

$$\frac{\partial n}{\partial x} = \frac{u_c(1 - t) + u_{cc}nx(1 - t)^2 - u_{cl}n(1 - t)}{-D}. \tag{9}$$

The sign of $\partial n/\partial x$ is indeterminate; as productivity increases, the supply of labor can be backward bending. Pretax income, $y = nx$, does not decline, however, even if n falls. Substituting (9) into (8) and rearranging terms shows that the bracketed term in (10) is the numerator of $\partial c/\partial r$ in (6). Hence, $\partial y/\partial x$ is positive for all $x > x_0$ provided that consumption is a normal good:

$$\frac{\partial y}{\partial x} = \frac{u_c(1 - t)x + n[u_{cl}x(1 - t) - u_{ll}]}{-D} > 0. \tag{10}$$

The final step in establishing that there is a unique equilibrium solution for any tax rate uses our assumption that leisure is a normal good. The government budget is balanced and all government spending is for redistribution of income. If per capita income is $\bar{y}$, then

$$t\bar{y} = r. \tag{11}$$

Let $F(\cdot)$ denote the distribution function for individual productivity, so that $F(x)$ is the fraction of the population with productivity less than x. Per capita income is obtained by integrating:

$$\bar{y} = \int_{x_0}^{\infty} xn[r, (1 - t)x]dF(x). \tag{12}$$

Equation (12) shows that per capita income, and therefore total earned income, is determined once we know x_0, t, and r. From (5), we know that x_0 depends only on t and r, and from (11) we know that, for any tax rate, there is at least one value of r that balances the budget.[7]

[7] The left side of (11) is nonnegative and is a continuous function of r that is bounded by $t\bar{x}$, where $\bar{x}$ is the average of x.

If leisure is a normal good, the value of r that satisfies (11) for each t is unique.[8]

Once r or t is chosen, the other is determined. The individual's choices of consumption and the distribution of his time between labor and leisure are determined also. The choice of r or t uniquely determines each individual's welfare and sets the size of government.

III. The Size of Government

The political process determines the share of national income taxed and redistributed. The many ways to make this choice range from dictatorship to unanimous consent, and each produces a different outcome. We call each political process that determines the tax rate a voting rule.

In this section, we consider any voting rule that allows a decisive individual to choose the tax rate. Two examples are dictatorship and universal suffrage with majority rule. A dictator is concerned about the effect of his decisions on the population's decisions to work and consume, but he alone makes the decision about the tax rate. Under majority rule, the voter with median income is decisive as we show below. We then show that changes in the voting rules and changes in productivity change the tax rate and the size of government.

The decisive voter chooses the tax rate that maximizes his utility. In making his choice, he is aware that his choice affects everyone's decision to work and consume. Increases in the tax rate have two effects. Each dollar of earned income raises more revenue but earned income declines; everyone chooses more leisure, and more people choose to subsist on redistribution. "High" and "low" tax rates have opposite effects on the choice of labor or leisure and, therefore, on earned income.

Formally, the individual is constrained to find a tax rate that balances the government budget, equation (11), and maximizes utility subject to his own budget constraint, equation (3). The first-order condition for the decisive voter is solved to find his preferred tax rate:

$$\bar{y} + t\frac{d\bar{y}}{dt} - y_d = 0, \tag{13}$$

where y_d is the income of the decisive voter.

[8] The normality of leisure means that $\partial l/\partial r > 0$ and, therefore, $\partial n/\partial r = -\partial l/\partial r < 0$. Since

$$\frac{\partial \bar{y}}{\partial r} = \int_{x_0}^{\infty} x\,\frac{\partial n}{\partial r}\,dF(x) < 0,$$

the left side of (11) is a strictly decreasing, continuous function of r. The right side of (11) strictly increases with r. This implies that there is a unique value of r that satisfies (11).

Roberts (1977) showed that if the ordering of individual incomes is independent of the choice of r and t, individual choice of the tax rate is inversely ordered by income. This implies that with universal suffrage the voter with median income is decisive, and the higher one's income, the lower the preferred tax rate. By making the additional assumption that consumption is a normal good, we have shown that incomes are ordered by productivity for all r and t. Combining Roberts's lemma 1 (1977, p. 334) with our results, we can order the choice of tax rate by the productivity of the decisive voter.[9] The higher an individual's productivity, the lower is his preferred tax rate.

Figure 1 illustrates the proposition and shows the effect on the tax rate of changing the voting rule. The negatively sloped line is the relation between individual productivity, x, and the individual's preferred tax rate. This line need not be linear.

The maximum tax rate, t_{max}, is chosen if the decisive voter does not work. An example is $x = x_{d1}$. In this case, $x \leq x_0$; the decisive voter consumes only r, so he chooses the tax rate (t_{max}) that maximizes r. Any higher tax rate reduces aggregate earned income, tax collections, and the amount available for redistribution. From equation (5), we see that the maximum tax rate must be less than $t = 1$.

As productivity rises from x_0 to $\bar{x}$, the tax rate declines from t_{max} to 0. At $x_d = \bar{x}$, the decisive voter is endowed with average productivity and cannot gain from lump-sum redistribution, so he votes for no redistribution by choosing $t = 0$.[10] From equation (5) and $u_r(0, \cdot) = \infty$, we see that everyone works when $r = 0$. If the decisive voter's productivity exceeds $\bar{x}$, t and r remain at zero and aggregate earned income remains at society's maximum.

Changes in the voting rule that spread the franchise up or down the productivity distribution change the decisive voter and raise or lower the tax rate. Our hypothesis implies that changing the position of the decisive voter in the distribution of productivity changes the size of government provided $x_0 < x_d < \bar{x}$. Major changes in x_d have occurred in two ways. Wealth and income requirements for voting were reduced or eliminated, gradually broadening the franchise and lowering the income of the decisive voter. Social security retirement systems grew in most countries after the franchise was extended. By increas-

[9] The formal statement of the result is: Consider any two pairs (r_1, t_1) and (r_2, t_2). If $t_2 > t_1$, then for all x: x is indifferent between (r_1, t_1) and (r_2, t_2) implies that x' weakly prefers (r_2, t_2) to (r_1, t_1) for all $x' < x$ and x'' weakly prefers (r_1, t_1) to (r_2, t_2) for all $x'' > x$; x strictly prefers (r_1, t_1) to (r_2, t_2) implies that x'' strictly prefers (r_1, t_1) to (r_2, t_2) for all $x'' > x$; x strictly prefers (r_2, t_2) to (r_1, t_1) implies that x' strictly prefers (r_2, t_2) to (r_1, t_1) for all $x' < x$. Note that this result does not require unimodality of voter preferences for tax rates.

[10] We have omitted public goods. In an earlier version we showed that under carefully specified conditions, public goods can be included without changing the result for redistribution.

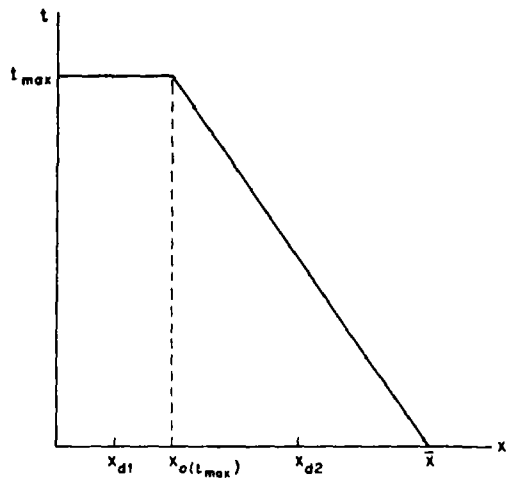


FIG. 1

ing the number of retired persons, social security systems increase the number of voters who favor increased redistribution financed by taxes on wages. Some of the retired who favor redistribution also favor low taxes on capital, property, and the income from capital.

The size of government changes also if there are changes in relative income, as shown by equation (13), or relative productivity. Conclusions about the precise effect of changes of this kind are difficult to draw. We cannot observe productivity directly and can only infer changes in the distribution of productivity, $F(\cdot)$, by observing changes in relative income. Recent literature makes clear that these effects are disputed (see Sahota 1978; King 1980; and others). Further, we cannot deduce the effect of changes in productivity on t directly from equation (13). The reason is that $\bar{y}$ depends on t, so finding the effect of changes in relative productivity requires the solution to a nonlinear equation in t. Instead, we rewrite (13) in a form which involves the (partial) elasticities of per capita income ($\bar{y}$) with respect to redistribution (r) and the wage rate ($x[1 - t]$).

Let $\tau = 1 - t$ be the fraction of earned income retained. From (12), $\bar{y}$ depends on r and τ only. The total derivative

$$\frac{d\bar{y}}{dt} = \frac{\bar{y}_r\bar{y} - \bar{y}_\tau}{1 - t\bar{y}_r}, \tag{14}$$

where $\bar{y}_r$ and $\bar{y}_\tau$ are the two partial derivatives. Substituting (14) into (13), we solve for t:

$$t = \frac{m - 1 + \eta(\bar{y}, r)}{m - 1 + \eta(\bar{y}, r) + m\eta(\bar{y}, \tau)}, \tag{15}$$

where m is the ratio of mean income to the income of the decisive voter, $\bar{y}/y_d$, and the η's are partial elasticities. Using the common economic assumption that the elasticities are constant, the tax rate rises as mean income rises relative to the income of the decisive voter, and taxes fall as m falls:

$$\frac{dt}{dm} = \frac{\eta(\bar{y}, \tau)[1 - \eta(\bar{y}, \tau)]}{[m - 1 + \eta(\bar{y}, \tau) + m\eta(\bar{y}, \tau)]^2} > 0. \tag{16}$$

Relaxing the assumption of constant elasticities weakens the conclusion, but we expect the sign of (16) to remain positive provided the change in the elasticities is small.

One of the oldest and most frequently tested explanations of the growth of government is known as Wagner's law. This law has been interpreted in two ways. The traditional interpretation is that government is a luxury good so that there is a positive relation between the relative size of government and the level of real income. Recently Alt (1980) has questioned this interpretation of Wagner's idea. Alt (1980, p. 4) notes that Wagner argued that there is "a proportion between public expenditure and national income which may not be permanently overstepped." This suggests an equilibrium relative size of government rather than an ever-growing government sector.

The traditional statement of Wagner's law—that government grows more rapidly than income—has been tested many times, but with mixed results. Peacock and Wiseman (1961), Cameron (1978), and Larkey et al. (1980) discuss these tests. Our hypothesis suggests that the results are ambiguous because Wagner's law is incomplete. The effect of absolute income on the size of government is conditional on relative income. Average or absolute income affects the elasticities in equation (15), and the relative income effect is given by m.

To make our hypothesis testable, we must identify the decisive voter. The applicable voting rule in the United States is universal franchise and majority rule. Under this rule, the voter with median income is decisive in single-issue elections, as we argued above. Hence the median voter is decisive in elections to choose the tax rate, so m is the ratio of mean to median income.[11]

[11] The many tests of the median-voter hypothesis using regression analyses are inconclusive. One reason is that many of the tests do not discriminate between the median and any other fractile of the income distribution (see Romer and Rosenthal 1979). Cooter and Helpman (1974) use income before and after taxes net of transfers to estimate the shape of the social welfare function implicit in U.S. data. They conclude that "the assumption that ability is distributed as wages per hour . . . —perhaps the best assumption on distribution of ability—vindicates the median voter rule."

IV. Conclusion

Government spending and taxes have grown relative to output in most countries with elected governments for the past 30 years or longer. Increased relative size of government appears to be independent of budget and tax systems, federal or national governments, the size of the bureaucracy, and other frequently mentioned institutional arrangements, although the relative rates of change in different countries may depend on these arrangements.

Our explanation of the size of government emphasizes voter demand for redistribution. Using a parsimonious, general equilibrium model in which the only government activities are redistribution and taxation, the real budget is balanced, and voters are fully informed, we show that the size of government is determined by the welfare-maximizing choice of a decisive individual. The decisive individual may be a dictator, absolute monarch, or marginal member of a junta.

With majority rule the voter with median income among the enfranchised citizens is decisive. Voters with income below the income of the decisive voter choose candidates who favor higher taxes and more redistribution; voters with income above the decisive voter desire lower taxes and less redistribution. The decisive voter chooses the tax share. When the mean income rises relative to the income of the decisive voter, taxes rise, and vice versa. The spread of the franchise in the nineteenth and twentieth centuries increased the number of voters with relatively low income. The position of the decisive voter shifted down the distribution of income, so tax rates rose. In recent years, the proportion of voters receiving social security has increased, raising the number of voters favoring taxes on wage and salary income to finance redistribution. A rational social security recipient with large property income supports taxes on labor income to finance redistribution but opposes taxes on income from property. In our analysis, there is neither capital nor taxes on property, so the increase in social security recipients has an effect similar to an extension of the franchise.

Our assumption that voters are fully informed about the size of government differs from much recent literature. There, taxpayers are portrayed as the prey sought by many predators who conspire to raise taxes relative to income by diffusing costs and concentrating benefits, or in other ways (Buchanan and Tullock 1962; Olson 1965; Niskanen 1971; Hayek 1979). We acknowledge that voters are ill informed about the costs of particular projects when, as is often the case, it is rational to avoid learning details. Knowledge of detail is not required to learn that the size of government has increased and that taxes have increased relative to output or income. Long ago it became rational for voters to anticipate this outcome of the political process.

Wagner's law, relating taxation to income, has generated a large literature and has been tested in various ways. Our analysis shows that Wagner's law should be amended to include the effect of relative income in addition to absolute income.

Kuznets (1955) observed that economic growth raises the incomes of skilled individuals relative to the incomes of the unskilled. In this way, economic growth can lead to rising inequality and, if our hypothesis is correct, to votes for redistribution. The rising relative size of government slows when the relative changes come to an end and reverses if the relative changes reverse in a mature stationary economy.

The distinctive feature of our analysis is not the voting rule but the relation between individual and collective choice. Each person chooses consumption and leisure by maximizing in the usual way. Anyone who works receives a wage equal to his marginal product. Taxes on labor income provide revenues for redistribution, however, so everyone benefits from decisions to work and incurs a cost when leisure increases.

The analysis explains why the size of government and the tax rate can remain constant yet be criticized by an overwhelming majority of citizens. The reason is that at the voting equilibrium nearly everyone prefers a different outcome. If unconstrained by the voting rule, everyone but the decisive voter would choose a different outcome. But only the decisive voter can assure a majority.

An extension of our argument may suggest why real government debt per capita, as measured in the budget, has increased more than 20-fold in this century. The decisive voter has as much incentive to tax the future rich as the current rich. An optimal distribution of the cost of redistribution would not tax only the current generation because, with economic growth, the future generation will be richer than the current generation. By shifting the burden of taxation toward the future, income is redistributed intertemporally.

To pursue these questions more fully and to analyze any effect of defense and public goods, it seems necessary to embed the analysis in a model with saving, capital accumulation, and public goods and to explore the effect of permitting relative shares to change as income changes. From an analysis of a growing economy, we can expect to develop a rational theory of the growth of government to complement our analysis of the government's size.

References

Alt, James. "Democracy and Public Expenditure." Multilithed. St. Louis: Washington Univ., 1980.

Aranson, Peter, and Ordeshook, Peter C. "Alternative Theories of the Growth of Government and Their Implications for Constitutional Tax and Spending Limits." Multilithed. Pittsburgh: Carnegie-Mellon Univ., 1980.

Browning, Edgar K., and Johnson, William R. *The Distribution of the Tax Burden.* Washington: American Enterprise Inst., 1979.

Brunner, Karl. "Reflections on the Political Economy of Government: The Persistent Growth of Government." *Schweizerische Zeitschrift Volkwirtschaft und Statis.* 114 (September 1978): 649–80.

Buchanan, James M., and Tullock, Gordon. *The Calculus of Consent: Logical Foundations of Constitutional Democracy.* Ann Arbor: Univ. Michigan Press, 1962.

Cameron, David. "The Expansion of the Public Economy: A Comparative Analysis." *American Polit. Sci. Rev.* 72 (December 1978): 1243–61.

Cooter, Robert, and Helpman, Elhanan. "Optimal Income Taxation for Transfer Payments under Different Social Welfare Criteria." *Q.J.E.* 88 (November 1974): 656–70.

Downs, Anthony. *An Economic Theory of Democracy.* New York: Harper & Row, 1957.

Hayek, Friedrich A. *Law, Legislation and Liberty.* Vol. 3, *The Political Order of a Free People.* Chicago: Univ. Chicago Press, 1979.

Hotelling, Harold. "Stability in Competition." *Econ. J.* 39 (March 1929): 41–57.

King, Mervyn A. "How Effective Have Fiscal Policies Been in Changing the Distribution of Income and Wealth?" *A.E.R. Papers and Proc.* 70 (May 1980): 72–76.

Kuznets, Simon. "Economic Growth and Income Inequality." *A.E.R.* 45 (March 1955): 1–28.

Larkey, P. D.; Stolp, C.; and Winer, M. "Theorizing about the Growth and Decline of Government: A Research Assessment." Multilithed. Pittsburgh: Carnegie-Mellon Univ., 1980.

Meltzer, Allan H., and Richard, Scott F. "Why Government Grows (and Grows) in a Democracy." *Public Interest* 52 (Summer 1978): 111–18.

Mueller, Dennis C. "Public Choice: A Survey." *J. Econ. Literature* 14 (June 1976): 395–433.

Musgrave, Richard A. *The Theory of Public Finance: Study in Public Economy.* New York: McGraw-Hill, 1959.

Niskanen, William A. *Bureaucracy and Representative Government.* Chicago: Aldine-Atherton, 1971.

Nutter, G. Warren. *Growth of Government in the West.* Washington: American Enterprise Inst., 1978.

Olson, Mancur, Jr. *The Logic of Collective Action: Public Goods and the Theory of Groups.* Cambridge, Mass.: Harvard Univ. Press, 1965.

Peacock, Alan T. *Economic Analysis of Government and Related Theories.* New York: St. Martin's, 1979.

Peacock, Alan T., and Wiseman, Jack. *The Growth of Public Expenditure in the United Kingdom.* Princeton, N.J.: Princeton Univ. Press, 1961.

Pechman, Joseph A., and Okner, Benjamin A. *Who Bears the Tax Burden?* Washington: Brookings Inst., 1974.

Peltzman, Sam. "The Growth of Government." Multilithed. Chicago: Univ. Chicago, 1979.

Riker, William H., and Ordeshook, Peter C. *An Introduction to Positive Political Theory.* Englewood Cliffs, N.J.: Prentice-Hall, 1973.

Roberts, Kevin W. S. "Voting over Income Tax Schedules." *J. Public Econ.* 8 (December 1977): 329–40.

Romer, Thomas. "Individual Welfare, Majority Voting, and the Properties of a Linear Income Tax." *J. Public Econ.* 4 (February 1975): 163–85.

Romer, Thomas, and Rosenthal, Howard. "The Elusive Median Voter." *J. Public Econ.* 12 (October 1979): 143–70.

Sahota, Gian Singh. "Theories of Personal Income Distribution: A Survey." *J. Econ. Literature* 16 (March 1978): 1–55.

Sheshinski, Eytan. "The Optimal Linear Income Tax." *Rev. Econ. Studies* 39 (July 1972): 297–302.

de Tocqueville, Alexis. *Democracy in America.* 1835. Reprint ed. Oxford: Oxford Univ. Press, 1965.

Wicksell, Knut. "A New Principle of Just Taxation." Reprinted in *Classics in the Theory of Public Finance,* edited by Richard A. Musgrave and Alan T. Peacock. New York: St. Martin's, 1958.

The Family, Inheritance, and the Intergenerational Transmission of Inequality

Nigel Tomes

University of Western Ontario

Unequal inheritance of material wealth is commonly considered a major cause of inequality in consumption. However, theoretical models of the intergenerational transmission of inequality by Becker, Blinder, and Ishikawa imply that unequal inheritance may either increase or reduce consumption inequality. Differences in inherited wealth resulting from unequal parental incomes increase inequality in recipients' consumption. However, unequal bequests caused by differences among families in the endowed ability of children or the costs of producing human capital are equalizing. Empirical results confirm these predictions: The inheritance received by children is inversely related to both children's income and parental education. Thus bequests are "compensatory" in that (ceteris paribus) low-income children inherit more than their advantaged contemporaries.

> The ultimate difficulties . . . center around the problem of social continuity in a world where individuals are born naked, destitute, helpless, ignorant and untrained. . . . The fundamental fact about society . . . is that it is made up of individuals who are born and die and give

An earlier version of this paper, "Inheritance and the Intergenerational Transmission of Inequality. Theory and Empirical Results," was presented in workshops at the University of Chicago, Queen's University, and the University of Western Ontario. Workshop participants provided helpful comments. Dorothy Worth provided programming assistance. I would like to thank Marvin B. Sussman and Judith N. Cates for permission to use their data. John A. Brittain provided a copy of the data. The referee and an editor provided constructive comments. Responsibility for any errors is my own.

[*Journal of Political Economy*, 1981, vol. 89, no. 5]
© 1981 by The University of Chicago. 0022-3808/81/8905-0002$01.50

> place to others; and . . . that it is dependent upon the utilization of three great accumulating funds of inheritance from the past, material goods and appliances, knowledge and skill, and morale [which] . . . must in some manner be carried forward to new individuals devoid of all these things as older individuals pass out. . . . [T]he institutions of the private family and private property, inheritance and bequest and parental responsibility, afford one way for securing more or less tolerable results in grappling with this problem. [KNIGHT 1921, pp. 374–75]

I. Introduction

Inequality in economic status and the reproduction of inequality in succeeding generations—"social mobility"—continue to be objects of social concern and the subject of investigation by social scientists. It has long been suspected that the institution of the family plays a major role in the generation and transmission of inequality and may severely limit the ability of public policy to alter the existing social order (Knight 1921).

The role of the family has been examined in a large number of empirical studies conducted within the human capital framework (e.g., Parsons 1975*a*). In contrast to this literature, there exists a research tradition which has examined the role of inherited material wealth in producing and perpetuating inequality (Stamp 1926; Wedgwood 1929; Blinder 1973, 1976*a*; Harbury and Hitchens 1979; Menchik 1979).

Empirical research on the intergenerational transmission of inequality has, for the most part, been conducted within the context of these two mutually exclusive approaches. However, models of family behavior recently proposed by Becker (1974), Ishikawa (1975), and Blinder (1976*b*) strongly suggest the possibility of substitution between alternative modes of transfer and their dependence on a common set of ability and opportunity variables. Thus, the "ability" of children and the education of parents—important determinants of children's schooling—should also influence parent-child transfers of material wealth.

Recent theoretical models suggest that the family may not contribute as much to inequality as is commonly thought. Since parents are predicted to adjust the magnitude of intergenerational transfers in response to differences in the endowments of children, the institution

of the family permits a redistribution of resources which tends to equalize consumption, both between parents and children (Stiglitz 1978; Becker and Tomes 1979)[1] and among disadvantaged and advantaged siblings (Becker and Tomes 1976; Griliches 1979). Because of the equalizing role of the family, government redistribution measures may have little, zero, or even a perverse effect on the distribution of income and consumption (Barro 1974; Becker and Tomes 1979).

The purpose of this paper is to empirically test the altruism model of the intergenerational transmission of inequality proposed by Becker, Ishikawa, and Blinder. In this model, parental bequests of material wealth and human capital investments represent substitute forms of intergenerational transfer and are simultaneously determined. The present paper extends previous work in a number of directions. First, previous studies of inheritance (Menchik 1979; Adams 1980) have been concerned with estimating the income elasticity of bequests[2]—and hence the extent to which unequal parental incomes are responsible for inequality in subsequent generations. In contrast, this paper emphasizes the response of parental bequests to variations in children's incomes resulting from differences in child ability or in the ability of parents to produce human capital in their children. Since intergenerational transfers serve in part to redistribute resources among family members, when parental income is held constant, the amount inherited is predicted to be inversely related to the recipient's income. Thus, within parental income strata, bequests of material wealth are predicted to be *equalizing*. Further, if more educated parents are more efficient at producing learning or earning skills in their children, the progeny of more educated parents will receive higher lifetime incomes from human capital. As a result, the equalizing role of the family will also be reflected in an inverse relationship between inherited wealth and parental education.

Second, the choice of family size, a variable emphasized in simulation models (Pryor 1973) and empirical studies (Lebergott 1976;

[1] In the model of Becker and Tomes (1979), it can be shown that, if the marginal propensity of parents to transfer resources to children exceeds the inheritability of endowments, the relative inequality in the steady-state distribution of income (as measured by the coefficient of variation) is less when parents choose the optimum level of transfers than if there were no intergenerational transfers.

[2] A limitation of these studies is that estimates of the income (wealth) elasticity of bequests do not distinguish between alternative models of bequest behavior. Both altruistic models and models of "egoistic" life-cycle accumulation in the presence of uncertain lifetimes and imperfect capital (annuity) markets are consistent with a high income elasticity of bequests (see Tomes 1979).

Blomquist 1979), is treated as endogenous,[3] as suggested by the "new microeconomic" approach to fertility.

Third, the observation that 55–60 percent of individuals do not receive an inheritance of material wealth (Blinder 1973; Blomquist 1979) suggests that there are two categories of households: those making transfers exclusively in the form of human capital who are at a corner solution making no transfers of material wealth; and second, households making transfers in both forms. These two groups (regimes) are predicted to respond differently to variations in parental income and education (Ishikawa 1975; Blinder 1976*b*; Edwards and Grossman 1977). In the absence of data on material wealth transfers, previous investigators (Edwards and Grossman 1977; Leibowitz and Friedman 1979) were restricted to indirect tests of this model. The present paper reports direct empirical tests of the two-regime model.

Empirical tests of the model are performed using a unique data set which contains information on the inheritance, income, and schooling of children and the occupation, education, and family size of parents.

The empirical results offer strong support for the Becker, Blinder, and Ishikawa models of family intergenerational transfers. The most striking empirical finding is that the inheritance received by children is inversely related to the recipient's income. This result supports the prediction that, within parental income strata, inheritance is equalizing—low-income children are "compensated" in the form of larger bequests, and high-income children are "taxed" in that they receive smaller bequests. Alternative explanations for the inverse relationship between inheritance and recipient's income are that it is the result of (i) the work-disincentive effect of nonlabor income on earnings, (ii) the payment by donors for greater old-age care rendered by low-income children, or (iii) the reallocation among siblings of shares in a given total bequest. These alternative explanations are not supported by the data, leading us to accept the hypothesis implied by the Becker, Blinder, and Ishikawa models: differences in inherited material wealth perform a role of equalizing the distribution of resources between successive generations of the family.

The plan of the remainder of the paper is as follows: Section II briefly outlines a model of the simultaneous determination of inheritance, human capital investment, and family size which captures the essential features of the Becker, Blinder, and Ishikawa models of

[3] Blinder suggests that "the taste for bequests . . . would rise with the number of children" (1976*b*, p. 90). Easterlin (1976) presents a "bequest" model of fertility in which parents choose the number of children consistent with achieving a given target bequest to each child. In the model outlined below, the bequest per child is simultaneously determined together with parental fertility.

intergenerational transfer behavior by the altruistic family. In Section III the data are described, and the model is specified empirically. Empirical results are reported in Section IV. Finally, Section V presents a summary and conclusions.

II. A Model of Inheritance, Human Capital Investment, and Family Size

Let us abstract from uncertainty and life-cycle factors[4] and consider a static, certain world in which individuals, endowed with perfect foresight, live for two periods—initially as a "child" and subsequently as a "parent." Generations of the family are assumed to be interdependent and overlapping. Parental fertility and the transmission of resources to children are assumed to be motivated by altruism, in that the number of children and their per capita consumption enter the parent's utility function together with the parent's own consumption. This utility function may be written as

$$U = U(N, z_c, z_p), \tag{1}$$

where N is the number of children and z_c is the per capita consumption of children during their subsequent tenure as parents. For simplicity, per capita consumption is assumed to be the same for all N children within a given family. The quantity z_p represents the aggregate consumption of parents.[5]

This formulation captures the essence of the Becker, Blinder, and Ishikawa models[6] in which parents derive satisfaction from a gross measure of the well-being of children—their lifetime utility, income, or consumption. The donor's utility does not depend on the composition of recipient's consumption (i.e., there are no "merit" goods), on direct services (e.g., old-age care), or on the sources of recipient's lifetime income—the division between human wealth and financial wealth.

Individuals are assumed to enter the world with an endowment of income determined by such factors as genetic and cultural inheritance, government fiscal and social policy, exogenous economic

[4] This assumption rules out "contingent bequests" when life-cycle savings for future consumption become bequests due to the uncertain incidence of mortality. Bevan and Stiglitz (1978) and Tomes (1979) discuss such models. In the present model, all bequests are planned and are motivated by parental altruism.

[5] All households are assumed to be composed of two parents plus children. I do not distinguish between the consumption of individual parents.

[6] The differences in the formations of Becker, Ishikawa, and Blinder may be of importance for some issues, such as the wealth effects of bond-financed fiscal policy. Of greater importance in the present context are the aspects that these theoretical models have in common.

growth, and random "luck." For simplicity, I assume that all children within a given family possess equal endowments. Thus, endowed income may vary across families but not within families.[7]

Motivated by their altruism, parents may augment the consumption of their children either by investing in their human capital or via direct transfers of material wealth.[8] The relationship between the child's consumption and parent-child transfers is determined by the income-expenditure constraint of the child generation and the technology described by the income-generating function,

$$z_c = I_c = e + \beta x^{1-\eta} e^{\gamma} + a = e + k + a, \tag{2}$$

where $k = \beta x^{1-\eta} e^{\gamma}$ is the constant elasticity human capital production function, $0 < \gamma, \eta < 1$.

The left-hand side of (2) represents the child's income-expenditure constraint which indicates that, in the absence of transfers by the child to subsequent generations, consumption equals income.[9] The right-hand side of (2) defines income in terms of its proximate determinants: e represents the endowed income of the child; k, income from human capital; and a, the child's income from material wealth.[10] Financial asset markets are assumed perfect, so that the rate of return on material wealth is constant and independent of the individual's stocks of human and nonhuman capital.

In general, both market-purchased goods (e.g., tuition) and home-produced inputs (e.g., the time of parents) enter the production function for human capital. For simplicity, we assume that market-purchased goods and home-produced inputs are perfect substitutes, so that a single aggregate input (x) enters the production function. Investment in the human capital of a given individual is assumed subject to diminishing returns with respect to x (i.e., $1 > \eta > 0$). The return to (home and market) inputs in the production of children's earning skills is also permitted to depend on the child's

[7] Becker and Tomes (1976) consider differences in transfers to children within the same family when children differ in endowed ability. Evidence relating to intrafamily inequality in bequests of financial wealth is presented in Sec. IV.

[8] We assume that all parents are at the interior solution where net parent-child transfers are positive—i.e., altruistically motivated transfers are nonzero. Even with the secular growth of incomes over the generations, regression toward the mean in endowed income and reverse (child-parent) transfers effected by bond-financed fiscal policies and social security programs would tend to make positive private parent-child transfers optimal for at least some households. Given the observed prevalence of "forward" transfers in forms such as schooling, this assumption seems plausible.

[9] Becker and Tomes (1979) consider intergenerational transfers in the context of many generations.

[10] For simplicity, stocks of both human capital and material wealth are measured in units of the child's income. Note also that in general e would represent a vector of earning and learning abilities.

characteristics—in the form of the child's endowments. To the extent that the child's endowment represents the ability to learn, greater child endowments will be associated with greater income for given inputs (i.e., $\gamma > 0$).

Families are permitted to differ in their ability to produce income-earning skills in their children (Leibowitz 1974). The parameter β in equation (2) represents the level of parental efficiency, which is assumed to operate in a manner equivalent to technological progress, increasing the next generation's income for given inputs and child endowments.

In the absence of intrafamily public goods and fixed cost components, and given identical children, the full income constraint of parents can be written as[11]

$$I_p = z_p + pNx + p_a Na, \tag{3}$$

where parental consumption (z_p) is the numeraire, I_p is the full income of parents, p is the unit price of human capital inputs, p_a is the cost of increasing the asset income of each child by \$1.00,[12] and pNx and $p_a Na$ represent (respectively) the total expenditures by parents on the human capital and material wealth of the next generation.

The Commodity Demand Functions

The allocation problem of parents is to maximize utility (1) subject to (2) and (3). The first-order conditions for an interior solution are:

$$\frac{U_p}{\Psi} = 1 \equiv \pi_p, \tag{4}$$

$$\frac{U_c}{\Psi} = \frac{pN}{\beta x^{-\eta} e^{\gamma}(1 - \eta)} = p_a N \equiv \pi_c, \tag{5}$$

$$\frac{U_N}{\Psi} = px + p_a a = p\left(\frac{k}{\beta e^{\gamma}}\right)^{1/(1-\eta)} + p_a a \equiv \pi_N, \tag{6}$$

where Ψ is the marginal utility of parental income, U_i is the marginal utility of consumption by the ith generation, and U_N is the marginal utility of family size.

The marginal rate of substitution between each commodity and parental income is equated to the marginal cost defined by the first-order conditions (π_j for the jth commodity). The interaction of

[11] Since all children have equal endowments and the same income-generating function (2), human capital investment and material wealth transfers will be equal for all children within a given family.

[12] In the absence of estate and income taxes, $p_a = (1 + r)^{-1}$, where r equals the intergenerational rate of return on assets.

"quality and quantity" dimensions of choice is reflected in the fact that the marginal costs of heir's consumption and family size (π_c and π_N, respectively) depend not only on prices but also on the level of commodity outputs. Because of this interdependence, we predict an inverse relationship between family size and parent-child transfers.

The optimum scale and mode of intergenerational transfers are determined by the first-order conditions (4) and (5). We assume that the (increasing) marginal cost of human capital investment is initially below the (constant) marginal cost of asset transfers. If loans from outside the family or intrafamily loans from parents are available, all households would invest in human capital up to the point where the marginal costs of both forms of transfer were equalized. Under these circumstances, human capital investment would be independent of the "opportunities" defined by parental income and family size and would be determined solely by parental efficiency and the child's endowed ability.

Observation suggests that such equality of opportunity among households does not exist as regards human capital investment because of the absence of collateral and the high cost of enforcing implicit or explicit loan contracts. In the present model, this imperfection in the capital market is modeled in an extreme form by assuming that all human capital investment is financed internally within the family and neither intergenerational loans nor bequests of debt are permitted (i.e., $a \geq 0$). Under this financial liability constraint, all human capital investment constitutes an intergenerational transfer. In this context there will exist two categories of families—those making transfers on a scale sufficient to make positive asset transfers for whom the financial liability constraint is ineffective, and those households who would like to transfer debt to the next generation but for whom the financial liability constraint is binding.

For households for whom the financial liability constraint ($a \geq 0$) is binding, the first-order conditions (5) and (6) require modification to become

$$\frac{U_c}{\Psi} = \frac{pN}{\beta x^{-\eta} e^{\gamma}(1-\eta)} \equiv \pi_c < p_a N, \tag{5'}$$

$$\frac{U_N}{\Psi} = px = p\left(\frac{k}{\beta e^{\gamma}}\right)^{1/(1-\eta)} \equiv \pi_N. \tag{6'}$$

This latter category of households will make all transfers in the form of human capital, and material wealth transfers will be zero. Since, for these households, marginal transfers are in the form of human capital investment, the level of such transfers will depend on parental income and family size—determinants of opportunities—in addition to parental efficiency and child endowments.

For both categories of households, the solutions to the appropriate first-order conditions can be represented as a set of demand functions for family size and the consumption of both generations in terms of "family resources" and commodity marginal costs (π_i's) (Becker and Lewis 1973). The demand for the marginal form of transfer represents an excess demand function for children's consumption.

If the parental utility function (1) is homothetic CES with equal elasticities of substitution between all arguments, and if we assume that the next generation's consumption (in present value terms) represents no more than half of family resources,[13] the following sign predictions are obtained:

For $a > 0$:

$$a = \Psi[\overset{(+)}{I_p}, \overset{(?)}{\beta}, \overset{(-)}{k}, \overset{(-)}{N}], \tag{7a}$$

$$N = \delta[\overset{(+)}{I_p}, \overset{(?)}{\beta}, \overset{(?)}{k}, \overset{(-)}{a}]. \tag{7b}$$

For $a = 0$:

$$P(a > 0) = \theta[\overset{(+)}{I_p}, \overset{(?)}{\beta}, \overset{(-)}{k}, \overset{(-)}{N}], \tag{8a}$$

$$N = \mu[\overset{(+)}{I_p}, \overset{(+)}{\beta}, \overset{(-)}{k}], \tag{8b}$$

where equation (8a) predicts the probability of positive material wealth transfers.

The major predictions of the model can now be summarized.[14] When asset transfers are positive, greater parental income is predicted to increase both asset transfers and family size. Since the demand for material wealth transfers represents an excess demand for child consumption, the ratio of asset transfers to family size will increase with parental income, and further, the income elasticity of material wealth transfers is predicted to be decreasing (Becker and Tomes 1976). The inverse relationship between quality and quantity dimensions of choice is reflected in the negative signs on family size and asset transfers in equations (7a)–(7b).

Material wealth transfers are predicted to be inversely related to children's income from human capital, reflecting the substitution between the two modes of transfer. When parental efficiency is held

[13] This latter condition is sufficient to sign some of the comparative static results. This condition will hold if the rate of time/generation preference—reflecting both "impatience" and the fact that consumption on the part of one's heirs is not the same as own consumption—exceeds the rate of return net of the (endogenous) rate of population growth, i.e., $\delta > r - g > 0$, where $g = N - 1$.

[14] The comparative statics and predictions of the model are discussed in detail in an earlier version of this paper. See also Tomes 1981*a*, 1981*b*.

constant, variations in human capital reflect differences in child endowments. An increase in child endowments raises the next generation's income for given parental expenditures, which increases family resources and therefore the consumption of both generations. The increase in parental consumption is financed by reduced parental expenditures on children. Since expenditures on children's human capital are an increasing function of child endowments, this requires a more than offsetting reduction in material wealth transfers.[15] In this way, the reduction in intergenerational asset transfers redistributes resources away from children with high endowment incomes and reduces the inequality in consumption between members of the next generation, as compared with the case when transfers are independent of endowments.

Equations (7a) and (8a) indicate that the effect of parental efficiency on both the magnitude and probability of receiving positive material wealth transfers is ambiguous when the recipient's human capital is held constant. However, when the scale of human capital investment is permitted to adjust (i.e., k is not held constant), greater parental efficiency is predicted to reduce both the magnitude and probability of receiving asset transfers.[16] Greater parental efficiency in the production of children's human capital reduces the cost of this mode of transfer and increases family resources. This produces both a substitution effect, increasing human capital investment and reducing material wealth transfers, and an income effect, increasing parental consumption, financed by reduced asset transfers.

A second set of predictions concerns the determinants of human capital investment in the two regimes—families making positive and zero material wealth transfers. Assuming all variables are measured in

[15] In the simple case in which $\gamma = \eta$, i.e., the human capital production function (2) is CRTS in x and e, the elasticity of material wealth transfers with respect to human capital transfers: k (holding β, I_p, and N constant) is given by: Λ ($d \ln a/d \ln k$) $= -(b_a(1 - b_c) + b_k\{1 - b_c[1 + (1 - \eta)\sigma]\})$, where $\Lambda = b_a(1 - b_c\sigma) > 0$, the b_i's are expenditure shares in family resources $S = I_p + \pi_c I_c$: $b_c = \pi_c z_c/S$, $b_a = \pi_c a/S$, $b_c = \pi_c e/S$, and $b_k = pNx/(1 - \eta)S$. Since $0 < 1 - \eta < 1$ and the elasticity of substitution $\sigma < 1$ (by the second-order conditions), a sufficient condition for the above expression to be negative is that the share of children's consumption (in present value terms) in family resources, b_c, not exceed ½.

[16] In the case considered in the previous footnote, the elasticity of material wealth transfers with respect to parental efficiency β, when human capital transfers are permitted to vary (holding I_p, e, and N constant) is given by: $\Lambda(d \ln a/d \ln \beta) = -b_k\eta^{-1}\{1 - b_c[1 + (1 - \eta)\sigma]\}$. This expression is negative if $b_c \leq$ ½. Further, it can be shown that in the reduced-form equation in which N is permitted to vary (obtained by substituting eq. [9a] into [7a]–[7b] and solving for a), greater parental efficiency reduces asset transfers, provided that the returns to parental expenditures on human capital do not exceed ½ (Tomes 1981*a*). Heckman's estimates of the returns to scale in the production of human capital do not differ significantly from ½ (Heckman 1976*a*, table 3A).

logs and given the assumptions concerning technology and the utility function, these predictions can be summarized as:

For $a > 0$:

$$k = \gamma_0 + \overset{(=0)}{\gamma_1 I_p} + \overset{(+)}{\gamma_2 e} + \overset{(+)}{\gamma_3 \beta} + \overset{(=0)}{\gamma_4 N}. \tag{9a}$$

For $a = 0$:

$$k = \rho_0 + \overset{(+)}{\rho_1 I_p} + \overset{(?)}{\rho_2 e} + \overset{(+)}{\rho_3 \beta} + \overset{(-)}{\rho_4 N}, \tag{9b}$$

$$\rho_1 > \gamma_1 = 0, \quad \gamma_2 > \rho_2, \quad \gamma_3 > \rho_3, \quad \rho_4 < \gamma_4 = 0. \tag{9c}$$

Since the financial liability constraint is not binding on families making positive asset transfers, the scale of human capital investment is independent of opportunity variables—implying the zero restrictions in (9a). For households at a corner solution making zero material wealth transfers on whom the financial constraint is binding, marginal transfers—in human form—depend both on parental income and family size.

An additional testable implication is that human capital investment is more responsive to parental efficiency among households making positive asset transfers. This follows from the prediction that, for households making positive asset transfers, the substitution effect on human capital transfers resulting from greater parental efficiency exceeds the sum of income and substitution effects experienced by families making no material wealth transfers (see Edwards and Grossman 1977; Tomes 1978). The predictions summarized in equations (7)–(9) are subject to empirical test in Section IV.

III. Empirical Specification

When we move from the theory to the empirical specification of the model, two points require discussion. First, the theoretical model assumed both perfect foresight and certainty. Under these conditions, individuals face a single allocation problem at the beginning of their parent status and have neither the desire nor the need to revise their decisions in the course of their lifetimes. However, in the real world both certainty and perfect foresight are lacking. Therefore, based on their future expectations, parents determine their fertility, the level of human capital investment, and material wealth transfers, with the possibility of revision if realized outcomes diverge from their expectations. Given that in the typical family life cycle completed fertility is determined prior to the completion of investment in the human capital of children and that material wealth transfers may occur at any time until the decease of the parents, these decisions are subject to

revision to differing degrees. In addition, human capital investment may be completed before the child's income-earning ability becomes known to either generation, since that ability may be subject to stochastic influences such as the occurrence of ill health and disability. Given that the parents' fertility is predetermined and given the precommitment of human capital investment, no adjustment of these choice variables may be possible or desirable. However, the level of material wealth transfers may be adjusted in the light of such events. In the empirical specification of the model I view the dependent variables as jointly determined. However, uncertainty and the temporal ordering of events suggest the introduction of variables which measure the realization of events over the life cycle.

Second, the theory derives predictions concerning the total transfers of material wealth to the next generation which may take the form of gifts during the parental lifetime or bequests at the decease of parents. Data on gifts are notoriously difficult to obtain, and the data set I use does not contain this information. However, the omission of information on gifts presents less of a problem for the present study which employs data on small wealth leavers. In these data there is little tax incentive in favor of gifts over bequests.[17] In addition, the prospect of unanticipated changes in the lifetime incomes of children and parents creates an incentive to defer material wealth transfers until the parents' decease. The inheritance received by the child is, therefore, used as a measure of material wealth transfers.

The Data

The data used in this study derive from a 5 percent random sample of estates probated in the Cleveland, Ohio, area in 1964–65, yielding 659 estates (Sussman, Cates, and Smith 1970). Surviving kin and other heirs were interviewed, including 657 sons and daughters of the decedents. Information was obtained on the total estate, usual occupation, education, and other characteristics of the deceased and on the inheritance, income, education, and other characteristics of the surviving kin. The principal variables are briefly described here and defined, together with other variables, in table 1.

[17] The mean gross estate of decedents was $12,000, and less than 5 percent of decedents' estates in this sample exceeded $60,000 after deductions and were liable for Federal Estate Tax (Sussman et al. 1970, p. 188). Ohio had (in 1964–65) an inheritance tax which commenced at 1 percent on the inheritance of children in excess of $7,000—almost double the mean inheritance in the data I use. Information gleaned from Barlow, Brazer, and Morgan (1966) suggests that the omission of data on gifts may not be a serious problem in terms of omitted variable bias and that the income elasticity of gifts exceeds that of bequests (for details, see the earlier version of this paper). Menchik's study of large Connecticut estates supports this latter conclusion (Menchik 1979, table 4).

The inheritance received by the child is used as a measure of material wealth transfers. Theory suggests that inheritances will be zero for families where financial constraints are binding. In accordance with this prediction, the observed distribution of inheritances is truncated—41.5 percent of children (251 cases) received the lower-limit value of $250[18] from the decedent's estate. The natural specification of the inheritance equation for the complete sample of children is a Tobit scheme:

$$\begin{aligned} a &= \beta X - u \quad \text{if} \quad \beta X - u > .250, \\ a &= .250 \quad\quad \text{if} \quad \beta X - u \leq .250, \end{aligned} \tag{10}$$

TABLE 1

DEFINITION OF VARIABLES

SAMPLE OF ALL CHILDREN		
Theoretical Variable	Empirical Measure	Mnemonic
Material wealth transfers	Inheritance received by son/daughter ($000s)*	INHR
Human capital investment	Annual family income of son/daughter (000s)†	INCOMER+‡
	Years of schooling of son/daughter	SCHR
Family size	Total number of kin of decedent	TNKIN
Parental income	Constructed measure of decedent's "permanent" income (000s)§	INCOMED+
Parental efficiency	Years of schooling of decedent	SCHD
Additional parent characteristics	Race of decedent (white = 1, nonwhite = 0)	RACE
	Sex of decedent (female = 1, male = 0)	SEXD
	Age of decedent	AGED
	Age of decedent squared	AGEDSQ
	Interaction between age and schooling of decedent	AGED · SCHD
	Marital status of decedent at death: married (omitted category: widowed, divorced, or single)	MARRIEDD
	Religion of decedent: Jewish, Protestant, Catholic (omitted category: no religion, Eastern Orthodox)	JEWD, PROTD, CATHD
	Origin of decedent: birthplace in U.S., Western Europe (omitted category: birthplace Eastern Europe or other)	ORIGINUS, ORIGINWE
	Number of kin of decedent, other than children	OTHERKIN

[18] Since the original data on inheritances are coded in $1,000 intervals, we were unable to distinguish between small inheritances (< $500) and zero inheritances (see table 1, §). In the discussion below we use the term "zero inheritance" to refer to children receiving an inheritance equal to the lower-limit value coded as $250 and "positive inheritance" to refer to an inheritance in excess of $250.

TABLE 1 (*Continued*)

Theoretical Variable	Empirical Measure	Mnemonic
	SAMPLE OF ALL CHILDREN	
Additional recipient characteristics	Sex of recipient (female = 1, male = 0)	SEXR
	Age of recipient	AGER
	Age of recipient squared	AGERSQ
	Marital status of recipient: single, divorced, or separated (omitted category, married)	SINGLER DIV/SEPR
	Number of household members employed	NEMPLR
	Dummy variable for the presence of nonlabor income (1: positive nonlabor income; 0: zero nonlabor income)	NLABINC
	Sex of breadwinner in recipient's household	SEX
	Frequency of contact between the recipient and decedent‖	VISITS

* In the original data the child's inheritance is coded in $000 intervals. An inheritance of less than $500 is therefore coded as $250 (i.e., 250), following Brittain (1978, p. 42). Estimates of the model using a limit value of zero were virtually identical with those reported.

† The reported monthly family income, which was originally coded in categories, was recoded on an annual basis, using interval midpoints for the closed intervals and using the (estimated) mean of a Pareto distribution for the open-ended interval.

‡ ' designates a variable entered in natural logarithm.

§ The income of recipients INCOMER (for those reporting a positive family income) was regressed on a set of "permanent" characteristics *of recipients*. The estimated coefficients were (R^2 = .216, N = 608)

$$\begin{aligned} \text{INCOMER}' = {} & 3.915 - .325\text{SCHR} + .01\text{AGER} \cdot \text{SCHR} - .104\text{AGER} \\ & + .0002\text{AGERSQ} + .266\text{OCCR} + .302\text{NEMPLR} - .093\text{SEX} \\ & + .627\text{RACE} + .494\text{ORIGINWE} + .658\text{ORIGINUS} + .320\text{ORIGINEE} \\ & + .668\text{PROT} + .539\text{CATH} + 1.278\text{JEW}, \end{aligned}$$

where ORIGINEE = 1 if recipient's origin is Eastern Europe and 0 otherwise, OCCR is a seven-category occupation of breadwinner code (coded from 1 unskilled, to 7 executive). These coefficients, together with the corresponding characteristics of the decedent, were used to construct the income variable INCOMED (NEMPL was set equal to unity).

‖ The original data on the frequency of contact were coded in categories: daily, weekly, monthly, less often, varies. These categories were recoded in terms of the number of visits per year (÷ 100). The category "less often" was coded as .03, i.e., three visits per year. The category "varies" and missing values were coded at the mean of the nonmissing observations.

where X is the vector of right-hand-side variables and u is assumed normally distributed, independent of X, with zero expectation and variance σ^2. In addition, we also present estimates of the determinants of the probability that an heir will inherit an amount in excess of the lower-limit value.

The level of schooling is the most frequently used measure of human capital investment. However, this variable corresponds to an input measure of human capital investment rather than the output measure suggested by the theory, which is the income of the child attributable to parental transfers. Also, schooling represents a decision variable determined early in the life cycle, whereas income represents the return to skills and other sources of income realized later in the life cycle. Life-cycle considerations suggest that family size may depend on children's schooling, whereas the inheritance received by

the child may be more closely related to the child's realized income than the child's schooling. The empirical model incorporates these considerations and includes an equation relating the recipient's income to schooling and other characteristics.

The reported total number of kin of the decedent was used as a measure of family size. Since this measure includes kin in addition to children, the number of these more distant kin was included as a control variable.

No direct measure of the annual or lifetime income of the decedent is available. However, information on the "permanent characteristics" of the decedent is available. Given supplementary data on income and the same set of characteristics from another source, one can construct a predicted income variable for decedents. The subsample of the surviving kin of these estates was used for this purpose. First, the log of recipient's family income was regressed on his permanent characteristics; then the estimated coefficients, together with the corresponding permanent characteristics of the decedent, were used to construct a measure of the decedent's "permanent" income. The education of the decedent was used as a measure of parental efficiency. When parental income (and hence market efficiency) is held constant, a higher level of parental education is interpreted as implying a greater level of nonmarket efficiency. Additional characteristics of the decedent (age, sex, race, religion, origin [birthplace], and marital status) and of the recipient (age, sex, marital status, and number of individuals employed in recipient's family) were included. Some of these variables hold constant life-cycle factors (e.g., age), while others may be interpreted as crude measures of some dimensions of the child's endowment.

Several writers have proposed that parental bequests may represent a payment by parents in return for services, such as care during old age, provided by children[19] (e.g., Sussman et al. 1970; Parsons 1975*b*). If low-wage children possess a comparative advantage in the provision of such services, as seems plausible, an inverse relationship between inheritance and the recipient's income could be observed even in the absence of altruism. In order to account for this possibility and in an effort to test this alternative hypothesis, an additional variable—the frequency of contact between the recipient and the decedent (VISITS)—was included in the inheritance and income equations. If some children sacrifice market income in order to care for their aged

[19] E.g., Sussman et al. (1970) state: "Testators will their estates to designated children . . . according to their perception of their needs, emotional ties with them, and services exchanged among family members over the years." Children "generally accept the notion that the sibling who has rendered the greatest amount of service to an aged parent should receive a major portion of the inheritance" (p. 118).

parents in return for a larger inheritance, the VISITS variable should enter with a negative coefficient in the income equation and a positive coefficient in the inheritance equation.

The empirical model consists of four equations in which the dependent variables are the inheritance, income, and schooling of the child and the total number of kin of the decedent. Since the determinants of human capital investment are expected to differ according to whether financial constraints are binding, estimates of the complete model are presented separately for the two subsamples according to whether the recipient inherited in excess of the limit value or the minimum amount.

Finally, in accordance with the previous discussion, life-cycle factors suggest certain identifying restrictions. Thus the marital status of the decedent at death—the presence or absence of a surviving spouse—is assumed to influence only the inheritance and to have no independent effect on the remaining endogenous variables. Also, the marital status of the recipient (divorced or separated, married or single) enters the schooling, income, and inheritance equations but is excluded from the family size equation.

IV. Empirical Results

Table 2 reports estimates for the sample as a whole. Row 1 presents the Tobit estimate of the inheritance equation, while row 2 presents the corresponding two-stage least-squares (2SLS) estimate. The important qualitative results are the same, but since the latter does not take account of the truncation of the dependent variable, I shall concentrate on the Tobit estimate.

The important finding is that the amount inherited is inversely related to the recipient's income, with an elasticity of -0.92 when evaluated at the mean inheritance.[20] This result contrasts with the findings of previous studies using these data.[21] Before accepting this result as confirmation of the theoretical model, several alternative interpretations should be considered. First, since the right-hand-side

[20] The regression in row 1 omits parental education. When this variable is included, it is insignificant ($t = 0.628$), while the recipient's income is on the margin of significance at the 5 percent level ($t = 1.932$).

[21] Brittain (1978), in an analysis of these data designed to detect differences in inheritance by the sex of the recipient, reported that "'need' for bequests indicated by the prebequest economic status of the son and daughter was tested, but found to have no influence. This is . . . consistent with equal division among siblings" (p. 45, n. 47). However, Brittain's regression (pp. 42–45) included only the number of children and the gross estate of the decedent as regressors in addition to the recipient's sex. Adams (1980) also analyzed these same data to estimate the income elasticity of bequests, but since the income of children was used as a measure of the decedent's income, he did not separate the effects of child's income and parent's income on material wealth transfers.

TABLE 2

INHERITANCE AND INCOME REGRESSIONS

Dependent Variable	Endogenous Variables: INHR[a]	INCOMER[b,c]	SCHR[c]	TNKIN[c]	INCOMED[b]	SCHD	VISITS	Other Variables	$\mathscr{L}$
1. INHR	.	−3.905 (2.080)		−8.087 (3.586)	3 581 (3.890)		−1.495 (2.347)	SEXD,* AGED,* ORIGINUS,* ORIGINWE, JEWD,* CATHD,* MARRIEDD,* OTHERKIN,* SINGLER,** AGER	Tobit 150.40
2. INHR	..	−3 993 (2.657)	. .	−4 462 (2 654)	4 579 (3.058)	−1 004 (1 403)	−1.140 (2 286)	SEXD,* AGED,* ORIGINUS, ORIGINWE, JEWD,* CATHD, MARRIEDD,* OTHERKIN,* SINGLER, AGER	2SLS
3. INCOMER[b]	.021 (2 166)		.255 (4 286)		. .		−1.102 (2.366)	AGER,* AGERSQ,** NEMPLR,* NLABINC, SEX, RACE, JEWD,* PROTD,* ORIGINUS,* DIV/SEPR	2SLS
4. INCOMER[b]	.018 (1 644)	.	(.166) (1.194)	−.033 (385)	−.149 (.979)	.131 (2.072)	−1.192 (2.389)	AGER,* AGERSQ,** NEMPLR,* NLABINC, SEX, RACE,* JEWD,* PROTD,** ORIGINUS, DIV/SEPR	2SLS
5. P_b(INHR > $250)	.	− 029[d] (2 127)	.	−.069[d] (4.186)	− 056 (1 119)	064 (2 535)	005 (.336)	SEXD,* AGED,* ORIGINUS,* ORIGINWE, OTHERKIN, MARRIEDD,* RACE,* JEWD, SEX	Logit 202.14
6. INHR	. .	.	..	..	7 281 (4 698)	−1 962 (2 480)	.	SEX,* AGED, ORIGINUS, ORIGINWE, PROTD,* CATHD,** MARRIEDD,* OTHERKIN,* AGER,* AGERSQ,* NLABINC	Tobit 148.54

NOTES —Sample all sons and daughters (N = 605). Absolute value of asymptotic t-statistics reported in parentheses beneath each coefficient All regressions included constant terms. Instruments in addition to the variables indicated above were SEXR, AGEDSQ, and AGED SCHD (an age-schooling interaction term) 2SLS, two-stage least squares Tobit Tobit regression 251 observations occurred at the lower-limit value INHR = $250 Logit Logit regression dependent variable is coded 1 if INHR > $250 ($N$ = 354), 0 if INHR = $250 ($N$ = 251) The reported coefficients are the marginal effects on the right-hand-side variables on the probability of receiving an inheritance in excess of $250, i e , $\beta_i P(1 - P)$, where β_i is the coefficient of the logistic function and P is the mean probability of INHR > $250 (i e., P = 585) $\mathscr{L} = (-2)$ ln [likelihood ratio] distributed χ^2 with $k - 1$ degrees of freedom, where k is the number of regressors.

[a] Variable predicted from a Tobit regression on all exogenous variables

[b] Variable entered in natural logarithm

[c] Variable predicted from an OLS regression on all exogenous variables

[d] Actual value rather than predicted

* Significant at 5% confidence level for variables for which coefficients and t-statistics are not reported

** Significant at 10% confidence level for variables for which coefficients and t-statistics are not reported

variable is the recipient's income rather than wage rate, it may reflect the labor supply of family members. Since the receipt of an inheritance is predicted to increase the demand for leisure, the work-disincentive effects of inheritance also imply an inverse relationship between inheritance and family earnings[22] (Heckman 1976*a*). In order to evaluate this possibility, the inheritance and income equations were estimated simultaneously in rows 2 and 3, with the recipient's income permitted to depend on the amount inherited.[23] The inheritance enters the income equation with a significant positive coefficient, suggesting that the income variable (INCOMER) may represent a better measure of income gross of the inheritance rather than earnings (see n. 22). The income equation reported in row 3 constrains the family background variables (family size, decedent's income and schooling) to enter recursively via the recipient's schooling. This assumption is relaxed in row 4 to allow for the direct influence of these variables on recipient's income via lifetime gifts and other mechanisms. The conclusions are unchanged—inheritance continues to enter with a positive coefficient which is on the margin of significance at the 10 percent level. The available evidence does not support the work-disincentive hypothesis

Second, as mentioned previously, an inverse relationship between children's income and inheritance may also arise if some children sacrifice market earnings in order to provide care for their parents, in return for a larger bequest. In order to test this possibility, the VISITS variable was introduced into both income and inheritance equations. This variable enters the income equation with a significant negative coefficient, suggesting that the opportunity cost of child-parent interaction is positive. Surprisingly, the VISITS variable enters with a significant negative coefficient in the inheritance equation.[24]

[22] In a single-period labor/leisure model the percent change in earnings (E) in response to a 1 percent change in nonlabor income (V) is given by $(d \ln E/d \ln V) = [K_V/(1 - K_V)](S_Z\eta_Z - 1) = [-K_V/(1 - K_V)]S_l\eta_l < 0$ if $\eta_l > 0$, where $K_V = V/I$, $S_l = wl/F$, $S_Z = P_Z Z/F$, where F is full income, I money income, l leisure, and Z consumption, and η_j is the income elasticity of demand for good j. The percent change in money income resulting from a 1 percent change in V is $(d \ln I/d \ln V) = K_V S_Z \eta_Z = K_V(1 - S_l\eta_l) > 0$ if $\eta_Z > 0$.

[23] In the income equation reported in row 3, the inheritance variable was predicted by using a Tobit regression to take account of the truncation problem. The results do not differ substantially when this variable is entered in the usual manner, ignoring the truncation at the lower limit. The income equation also contains a dummy variable for the presence/absence of nonlabor income (NLABINC) and the number of household members employed (NEMPLR), in order to hold constant differences in labor force participation between households. The estimated coefficient on schooling in the income equation is "large" in comparison with estimates of the rate of return to schooling. However, the dependent variable is (the log of) family income rather than individual earnings. If both the education levels and earnings of spouses are positively correlated, this could explain the large marginal effect of schooling.

[24] This conclusion differs from that reached by Sussman et al. (1970) based on the

This result presents prima facie evidence against the pure "child-services" model of inheritance (i.e., with no altruism), since the decedent as a monopsony demander of child services would never bequeath an inheritance such that the recipient located on the backward-bending region of the child-services supply schedule.[25] We therefore conclude that the negative relationship between the amount inherited and the recipient's income confirms the existence of an operative altruistic bequest motive and is consistent with the prediction that inherited material wealth serves as a means of redistributing resources among family members.

Other findings may be briefly summarized. The decedent parent's income enters the Tobit inheritance equation (row 1) with a significant positive coefficient, implying an income elasticity of 0.84 when evaluated at the mean inheritance. Last, the total number of kin (TNKIN)—the empirical measure of family size—enters the inheritance regression with a significant negative coefficient, consistent with the existence of substitution between family size (quantity) and the per capita bequest (quality).

Row 5 reports a logit regression in which the dependent variable is the probability of receiving an inheritance in excess of the lower-limit value.[26] This equation represents a test of the ability of the model to

study of various cases: "The data indicate general agreement that the child who performs the greatest amount of service should receive the greatest part of the estate" (p. 119). The results reported here do not invalidate this explanation of estate distribution in specific cases. Post hoc explanations of this finding are (i) that contact with aged parents is a time-intensive and inferior commodity or (ii) that children reporting the greatest frequency of contact are compensated during the parental lifetime (e.g., by living rent free in the parental home). A more complete model would include VISITS as an endogenous variable. Along these lines a simple OLS "supply of visits" equation was estimated (t-statistics in parentheses; † = ln):

$$\begin{aligned}
\text{VISITS} = {} & 2.593 - \underset{(2.831)}{.097\text{INCOMER}^{\dagger}} - \underset{(1.021)}{.004\text{INHR}} + \underset{(2.743)}{.356\text{NLABINC}} \\
& + \underset{(3.950)}{.790\text{SINGLER}} + \underset{(4.786)}{.560\text{SEXR}} + \underset{(3.528)}{.314\text{NEMPLR}} \\
& - \underset{(.078)}{.008\text{INCOMED}^{\dagger}} - \underset{(.481)}{.025\text{SCHD}} - \underset{(4.059)}{.138\text{TNKIN}} \\
& - \underset{(1.670)}{.561\text{RACE}} - \underset{(2.169)}{.282\text{MARRIEDD}} + \underset{(2.456)}{.700\text{JEWD}};
\end{aligned}$$

$$R^2 = .162, \qquad N = 605.$$

The recipient's income enters with a significant negative coefficient, while the inheritance coefficient is negative but not significant. These results are subject to possible simultaneous-equations bias.

[25] However, this finding is consistent with a child-service model which incorporates parental altruism.

[26] The logit procedure assumes the distribution of u is Sech^2 rather than normal. However, the cumulative distribution of the logistic function closely approximates the normal except in the tails (Nerlove and Press 1973). As expected, a probit regression produced similar results.

assign individuals to the alternative regimes[27] (see eq. [8a]). As predicted by the theory, the recipient's income and the number of kin enter with significant negative coefficients.[28] In contrast to the Tobit results (row 1), neither the decedent's income nor the VISITS variable is significant in the logit regression, while the decedent's schooling enters with a significant positive coefficient.

The final row of table 2 reports a reduced-form Tobit inheritance equation. This regression allows us to determine the effects of parental income and education on inheritance when family size and the recipient's income are permitted to vary. In this equation, the income elasticity of inheritance is 1.71, evaluated at the mean of the dependent variable—a value significantly different from unity at the 6 percent level. This estimate is considerably larger than the estimate in row 1, a result consistent with the prediction of Becker and Tomes (1976) that, because of the effects of quality-quantity interaction, the "observed" income elasticity of child quality will exceed the underlying "true" elasticity. The decedent's schooling enters the reduced-form inheritance regression with a significant negative coefficient. This confirms the prediction that more educated parents invest more in their children's human capital (see table 3) and as a consequence make smaller bequests of material wealth.[29] The parameter estimate implies that each additional year of decedent's schooling reduces the child's inheritance by almost $2,000—a sizable effect in relation to the average inheritance of $4,257.

Estimates of the Complete Model

Table 3 presents estimates of the four-equation model. In these regressions the sample is partitioned into two categories according to whether children inherit in excess of the limit value or not. Since the self-selection of households into either category results in the possibility of sample-selection bias, the correction procedure outlined by Heckman (1976*b*, 1979) was implemented. This procedure involves initially computing Mills ratios from the probit equation of the selection rule and including these "omitted" variables—designated "λ_a"

[27] As an intuitive test of ability of the logit equation to assign individuals to regimes, recipients were assigned to 0,1 categories according to the predicted $\hat{P} \lesseqgtr .5$. This resulted in the "correct" assignment of 73.12 percent of recipients as compared with a 50 percent "correct" assignment based on random assignments using the population proportions.

[28] This first result confirms the finding of Sussman et al. (1970) that "those most likely to inherit were persons with the smallest monthly incomes" (p. 155). However, those authors also state that "the percentage of survivors who inherited did not vary greatly among [survivor] income categories" (ibid.).

[29] Additional empirical evidence in support of this prediction is reported in Tomes (1981*b*), where it is found that the total estate of the decedent is inversely related to the decedent's education, holding the decedent's income and other characteristics constant.

TABLE 3

ESTIMATES OF THE MODEL

Dependent Variable	Endogenous Variables								Other Variables
	INHR	INCOMER[a]	SCHR	TNKIN	INCOMED[a]	SCHD	VISITS	λ_a	
	Children Receiving an Inheritance in Excess of the Limit Value (N = 354), λ_a								
1. INHR	...	−5 233 (3.228)	...	−6.499 (4.516)	5.847 (3.202)	−2.289 (2.144)	−.499 (.830)	9.572 (1.442)	AGED,* MARRIEDD, JEWD,** OTHERKIN,* SINGLER
2. INCOMER[a]	−.009 (.453)	...	.292 (3.443)	...	...	...	−.098 (1.540)	.385 (1.452)	AGER, AGERSQ, NEMPLR,* NLABINC, DIV/SEPR, SEX, RACE, JEWD**
3. SCHR	...	...	...	−.335 (1.508)	.281 (1.355)	.249 (2.220)	...	−.231 (.536)	OTHERKIN, RACE, ORIGINUS, ORIGINWE, JEWD,* SINGLER,* AGER,* AGED*
4. TNKIN	−.056 (4.031)	..	−.208 (2.336)	...	.448 (2.484)	−.160 (1.771)	...	.718 (2.453)	OTHERKIN,* RACE,** CATHD,* JEWD, AGED,* ORIGINUS,* AGER*
5 INHR[a]	...	− 197 (2.042)	..	−.294 (3.625)	.363 (3.764)	−.142 (2.468)	−.041 (1.155)	.891 (2.311)	AGED,* MARRIEDD, JEWD, OTHERKIN,* SINGLER
6. INCOMER[a]	−.054[a] (.148)	...	.259 (2.768)	...	...	...	−.111 (1.721)	.432 (.990)	AGER, AGERSQ, NEMPLR,* NLABINC, DIV/SEPR, SEX, RACE, JEWD**

7. SCHR	...	..	...	−.360 (1.627)	286 (1 377)	.251 (2.231)	...	−.234 (.543)	OTHERKIN, RACE, ORIGINUS, ORIGINWE, JEWD,* AGED,* SINGLER,* AGER*
8. TNKIN	−.457[a] (1.459)	..	−.177 (1.716)	...	.264 (1.405)	−.108 (1.024)	...	1.049 (2.157)	OTHERKIN,* RACE,** CATHD,* JEWD, ORIGINUS,* AGED,* AGER
	Children Receiving a Zero Inheritance (N = 251), λ_b								
9. INCOMER[a]	...	..	.415 (5.553)	...	...	..	−.134 (2.008)	.670 (2.127)	AGER,* AGERSQ,** NEMPLR, NLABINC,* SEX,** RACE, JEWD
10. SCHR	...	..	..	−.273 (1.060)	.791 (2.232)	.077 (.423)	...	.068 (.123)	OTHERKIN, RACE, ORIGINUS,** ORIGINWE, JEWD, SINGLER,* AGER,* AGED
11. TNKIN	...	..	−.252 (2.464)	...	−.262 (1.297)	.258 (2.681)	...	.007 (.019)	OTHERKIN,* RACE,* PROTD, CATHD,* ORIGINUS, AGED,* AGER

NOTE.—All equations estimated by three-stage least squares. Asymptotic t-statistics reported in parentheses beneath coefficients. λ_a, λ_b, Mills ratios computed from a probit regression on all exogenous variables (see text). See notes to table 2.

* Variable entered in natural logarithm.

and "λ_b" in table 3—in subsequent regressions using the selected sample.[30] Table 3 reports the results obtained using 3SLS, which allows the error terms of the equations to be correlated.

Rows 1–8 report results for children who received a bequest in excess of the limit value. The estimates of the inheritance and income equations duplicate quite closely those obtained for the whole sample (table 2, rows 1–3) and require little additional discussion. Recipient's income and parents' income enter the inheritance equation with the predicted negative and positive signs, with elasticities (at the mean) of -0.74 and $+0.82$, respectively.[31] In contrast to the results in table 2, inheritance is not significant in the income equation (row 2), and the VISITS variable is not significant in line 1 or 2.[32]

Although the semilog specification of the inheritance equation may be preferred on the a priori grounds that it embodies the prediction of a decreasing income elasticity, the constant elasticity formulation was also estimated (row 5). The estimates show that the major qualitative results are not dependent on the use of the semilog functional form.

The results reported in table 3 represent a direct test of the two-regime model. The major question to be addressed is whether the

[30] If eq. (10) represents the inheritance equation, the expected values of variables Z_j for the censored samples can be shown (see Willis and Rosen 1979) to be

$$E(Z_j \mid a > .250) = \gamma_j X + E(\epsilon_j \mid \beta X > u) = \gamma_j X + \frac{\sigma_{ju}}{\sigma_u}\lambda_a,$$

$$E(Z_j \mid a = .250) = \gamma_j X + E(\epsilon_j \mid \beta X \leq u) = \gamma_j X + \frac{\sigma_{ju}}{\sigma_u}\lambda_b,$$

where

$$\lambda_a \equiv E\left(\frac{u}{\sigma_u} \middle| \frac{u}{\sigma_u} < \frac{\beta X}{\sigma_u}\right) = -f(\beta X/\sigma_u)/F(\beta X/\sigma_u) \leq 0,$$

$$\lambda_b \equiv E\left(\frac{u}{\sigma_u} \middle| \frac{u}{\sigma_u} \geq \frac{\beta X}{\sigma_u}\right) = f(\beta X/\sigma_u)/[1 - F(\beta X/\sigma_u)] \geq 0,$$

where F is the cumulative normal density and f is its probability density function. The estimation procedure employs the estimated values $\hat{\lambda}_a$ and $\hat{\lambda}_b$ from a probit regression including all the exogenous variables except VISITS—because of the dubious exogeneity of this variable (see n. 24). The presence of the selection rule introduces heteroscedasticity into the equation residuals, resulting in upward-biased t-statistics. However, in the present context, there is a presumption that the bias is small. Applying the appropriate correction procedure to the t-statistics from the 2SLS estimates resulted in only marginal changes in the t-statistics.

[31] The significant positive coefficient on λ_b in row 9 indicates that individuals who inherit zero receive higher incomes than individuals with the same income-earning characteristics in the whole sample. This also supports the compensation hypothesis.

[32] The income equation (row 2) constrains the background variables (family size, decedent's income and schooling) to enter recursively via the recipient's schooling. When these variables are entered directly (results not reported here), the major conclusions are unaffected.

determinants of human capital investment differ among households according to whether material wealth transfers are positive or zero. A comparison of the schooling regressions for the two subsamples (rows 3 and 10) indicates differences which are in the predicted direction. For children receiving an inheritance in excess of the limit value, the level of schooling is unrelated to either parental income or family size but is positively related to parental education. In contrast, for children who received a zero inheritance, the level of schooling depends on parental income (with an income elasticity of 0.07) but is unrelated to parental education. Thus opportunities (parental income) are important for households making zero material wealth transfers, while parental education (one determinant of ability) is more important for households making positive material wealth transfers. However, contrary to the theoretical prediction, the family size variable—a determinant of opportunities—is not significant in the school equation for children who inherit zero (row 10) upon whom financial constraints are expected to be binding.

The family size regressions also differ between the two subsamples (rows 4 and 11), although concerning these differences we have no predictions. For families making asset transfers in excess of the limit value (row 4), parental income has a positive and significant coefficient—the only evidence in these data of a positive income elasticity of family size. Parental education enters with a negative coefficient which is significant at the 10 percent level. In contrast, for the sample receiving a zero inheritance (row 11), the parental income variable is negative and not significant, while parental education enters with a significant positive coefficient. These results underline the complexity and ambiguity of the relationships among family size and parental income and education. The one result common to both subsamples is that measures of child quality (the child's inheritance and/or schooling) enter with significant negative coefficients in the family size equations, consistent with the existence of a quality-quantity trade-off.[33]

Additional Results

A major finding of the foregoing analysis is that, when parental income and education are held constant, children with higher incomes receive a smaller inheritance of material wealth. This is consistent with the theoretical prediction that parents respond to dif-

[33] The significant positive coefficients on λ_a in rows 4 and 8 imply that those who receive a positive inheritance are more likely to come from a smaller family than random individuals with the same characteristics. This also implies a quality/quantity trade-off.

ferences in the ability of children by reducing transfers to childr order to finance greater parental consumption. In this way, pa share in the returns to greater child endowments. However, an native interpretation is that the observed inverse relationship bet inheritance and children's income reflects (partly or entirely) the of inheritance in compensating for differences *among* siblings r than between parents and children. Thus, if parents beque larger share of the estate to the child with the lowest income smaller share to his advantaged sibling, this also implies an in relationship between inheritance and the recipient's income.

This issue can be addressed using data on 137 families for w information is available on more than one child. In this subsar 41.6 percent of heirs (siblings) received exactly equal shares. ever, this figure includes cases in which all heirs received the value of zero and therefore, by definition, inherited equally. Ex ing such cases, 21.1 percent of heirs inherited exactly equal sh and the vast majority of heirs (78.9 percent) received unequal s in the parental estate. However, some of these between-siblin ferences are small: 50.4 percent of heirs received approximately shares—within $500 of the average inheritance received by chil Nevertheless, the incidence of equal division is dramatically l than that found by Menchik (1980) in his sample of large Conne estates. Menchik found that "equal sharing among children i rule" (p. 299). For example, among two-child families, in 62.5 pe of cases both children received exactly equal bequests and in percent of cases approximately equal bequests (Menchik 198 310). The corresponding figures for two-child families in the pr sample are 22.2 percent and 44.4 percent.[34] The fact that st equal division is *not* the rule in this sample suggests that the possi that inheritance is equalizing between siblings warrants investiga

Table 4 reports the results of Tobit regressions for various sub ples of children from families with information on multiple heirs availability of multiple observations per family also permits the i sion of an unobserved family component in the income predi equation. The constructed measure of the decedent's income

[34] Two possible explanations of this difference are: (i) In large estates donor make sizable transfers in the form of inter vivos gifts in order to avoid estat inheritance taxes. Thus, unequal lifetime gifts by wealthy donors may serve the functions as unequal bequests by less wealthy donors. (ii) Apparent equal divis probate records is consistent with de facto unequal division when estates are rec uted among heirs. Sussman et al. (1970) report that, in 27 percent of intestate (i.e., the decedent left no will) for which the law prescribed equal division siblings, ex post redistribution among siblings occurred resulting in unequal shar 143–44). Thus probate records may underestimate the incidence of unequal div

TABLE 4

ADDITIONAL RESULTS FOR FAMILIES WITH MULTIPLE HEIRS

DEPENDENT VARIABLE	$\overline{\text{INCOMER}}$[a,b]	DINCOMER[a,c]	TNKIN[b]	INCOMED[a]	SCHD	OTHER VARIABLES	$\mathscr{L}$	N	N Limits
1. INHR	−8 053 (3.009)	−1.077 (1.467)	−1.855 (1.695)	6.870 (4.536)	.887 (1.266)	SEXD,* ORIGINUS, ORIGINWE, JEWD,* MARRIEDD, $\overline{\text{SEX}}$,* AGER*	71.85	250	60
2. INHR	−9.196 (2.678)	−1.237 (1.377)	−2.113 (1.373)	8.211 (4.612)	784 (.905)	SEXD,* ORIGINUS, ORIGINWE, JEWD, MARRIEDD, $\overline{\text{SEX}}$,* AGER*	64.91	196	60
3. INHR	−14.092 (2.882)	−1 998 (1.611)	−4.056 (2.109)	9.652 (3.509)	1.524 (1.168)	SEXD,* ORIGINUS, ORIGINWE, JEWD,* MARRIEDD, $\overline{\text{SEX}}$, AGER*	46.08	119	23

NOTE — Tobit regressions, asymptotic t-statistics reported in parentheses, $\mathscr{L}$ = (−2) ln (likelihood ratio), N limits = limit observations (INHR = $250). $\overline{\text{INCOMER}}$* is the mean predicted (ln) income of children in the same family. DINCOMER is the deviation of the child's income from the mean income of children (i.e., DINCOMER* = INCOMER* − $\overline{\text{INCOMER}}$*). $\overline{\text{SEX}}$ is the mean sex of the family breadwinner in the siblings families. The INCOMED* variable in these regressions contains an unobserved family component. The weight attached to this unobserved component in constructing INCOMED was determined by minimizing the standard error of the regression. The weights used in the reported regressions in rows 1–3 were .5, .7, and .5, respectively (see text for details). Samples: (i) row 1, multiple heirs where at least one child received an inheritance in excess of the limit value, (ii) row 2, same as (i), excluding cases where siblings inherited exactly equal shares, (iii) row 3, same as (i), excluding cases where siblings inherited within $500 of the average inheritance of children (i.e., cases of approximately equal division). See table 2 notes.

[a] Variable entered in natural logarithm.

[b] Predicted value of variable from OLS regression on all exogenous variables, using whole sample (N = 605).

[c] Actual rather than predicted value.

ployed in table 4 includes this family component.[35] In these regressions, both the mean income of children and the deviation of the child's income from the family mean are introduced as determinants of inheritance. The latter variable is predicted to enter with a negative coefficient if disadvantaged children receive larger shares in the parental estate. This variable enters with the predicted negative sign but never attains statistical significance. In contrast, the mean income of children enters consistently with a highly significant negative sign. Estimates of the elasticity of inheritance with respect to the mean income of children cluster closely around −1.8, evaluated at the mean of the dependent variable. Thus, when we hold income inequality among children constant, the higher the average income of children, the lower the inheritance received by children. These results suggest that the inverse relationship between recipient's income and inheritance reflects for the most part a reallocation of resources between the generations of the family rather than simply the redistribution among siblings of a given total bequest to children.

These results for the subsample for which information is available on multiple heirs and also those for the subsample with information on both the decedent and surviving spouse[36] (not reported here) confirm the results reported earlier for the whole sample.

V. Summary and Conclusions

I have presented an empirical model of the intergenerational transmission of inequality within the altruistic family. This paper adds

[35] Adams (1980) employed a conceptually similar, but less appealing, technique using the mean income of recipients in a family to measure the decedent's income. The procedure employed here was to estimate an income prediction equation similar to that in table 1, §, for the subsample for whom information is available for two or more survivors—either children or the surviving spouse (N = 431). The mean residual was then computed for each family and entered in a subsequent income equation. The estimated coefficients from this second regression, together with the decedent's characteristics and the family mean residual, were used to construct the measure of the decedent's income. The weight attached to the family residual depends on the extent to which the unobserved family component is correlated between spouses and transmitted from parents to children. In the absence of information on this parameter, weights were used in the unit interval (0.0, 0.1, . . . , 1.0). The reported results employ the weight which minimized the standard error of the regression (see note to table 4). In practice, the measure of the decedent's income used in tables 2–3 performed almost as well as the measure used in table 4. (See the results reported in the earlier version of this paper.) The estimated income elasticity of inheritance did not differ in a systematic fashion using the two alternative income measures.

[36] For this subsample (N = 191), the effects of mother's and father's education, ages, and incomes were analyzed; also, a measure of the predicted bequest from the surviving spouse to children was constructed and used in the subsequent regressions. The results, which confirm the basic conclusions of this study, were reported in the earlier version of this paper.

empirical content to the theoretical models of Becker, Blinder, and Ishikawa in which bequests of material wealth and investment in children's human capital are simultaneously determined and represent substitute modes of parent-child transfers.

Empirical tests of the model were performed using a unique data set containing information on the inheritance, income, and schooling of the surviving kin of a random sample of estates probated in the Cleveland, Ohio, area in 1964–65.

The empirical results strongly confirm the equalizing role of inheritance and the existence of substitution between human capital investment and inherited material wealth. The inheritance received by the child was found to be inversely related to the child's income. Material wealth transfers are therefore "compensatory"—in that (other things being equal) children with low incomes receive greater bequests of material wealth than their better-endowed contemporaries. Our estimates imply that, for individuals in the sample who received a positive inheritance (table 3, row 1), the compensatory component of bequests reduced the variance of inherited wealth by 30 percent and reduced the correlation between inheritance and income from 0.53 to 0.12. This result confirms the prediction of the altruism model of intergenerational transfers that an increase in the income of recipients leads the donor to reduce transfers in order to redistribute family resources in favor of his own consumption. The data do *not* support the alternative explanations of this inverse relationship—that the receipt of an inheritance has a work-disincentive effect, that low-income children inherit more in return for the greater child services they provide for their aged parents, or that the observed relationship reflects the reallocation among siblings of shares in a given parental estate. Differences across families in parental education were found to result in variations in inherited wealth. Greater parental education is associated with more schooling and higher incomes of children and lower bequests of material wealth—results consistent with the theoretical predictions of the altruism model.

There is evidence that the financial constraints faced by families differ across households. Families who make positive material wealth transfers appear less constrained financially when it comes to investing in their children's human capital—and ability variables are the major determinants of children's schooling. Conversely, families making zero bequests to children appear more constrained financially, so that opportunity variables are the major determinants of children's schooling. For both categories of households, there is evidence of a trade-off between family size (quantity) and per capita parent-child expenditures (child quality).

The first result is important for a number of reasons. Previous empirical work (Brittain 1978; Menchik 1980) has failed to find any evidence of compensatory bequests among siblings. In contrast, the present study provides strong evidence that within the family—between parents and children—and among families in the same income/education stratum, the distribution of inherited wealth is equalizing. This supports the recent suggestion of Griliches (1979), that the family contributes less to the intergenerational transmission of inequality than is commonly supposed, since it operates as a force for equality between successive generations (parents and children).

Appendix

MEANS OF VARIABLES FOR VARIOUS SAMPLES

Variable	All Sons/Daughters Reporting Positive Family Income (N = 605) (1)	Sons/Daughters in col. 1 Receiving an Inheritance in Excess of the Limit Value (N = 354) (2)	Sons/Daughters in col. 1 Receiving an Inheritance Equal to the Limit Value (N = 251) (3)
INHR ($000s)	4.257	7.099	.250
INCOMER ($000s)	11.518	11.922	10.973
SCHR	12.570	12.517	12.645
AGER	42.828	44.384	40.633
SEXR (% female)	49.3	48.6	50.2
NEMPLR	1.190	1.184	1.199
NLABINC (% reporting the receipt of nonlabor income)	29.8	31.4	27.5
SINGLER (%)	9.6	9.0	10.4
DIV/SEPR (%)	.3	.3	.4
INCOMED ($000s)	6.085	5.559	6.906
SCHD	9.102	9.054	9.171
AGED	70.630	72.056	68.618
SEXD (% female)	41.7	52.5	26.3
MARRIEDD (%)	58.8	39.5	86.1
TNKIN	4.134	3.791	4.618
OTHERKIN	1.020	.743	1.410
RACE (% nonwhite)	3.3	2.7	5.6
CATHD (%)	44.1	44.9	43.0
PROTD (%)	43.6	41.5	46.6
JEWD (%)	5.1	6.2	3.6
ORIGINUS (%)	51.6	45.2	60.6
ORIGINWE (%)	16.2	18.1	13.5
VISITS	1.443	1.550	1.292
λ_a	−.696	−.506	−.965
λ_b	1.068	1.319	.713

References

Adams, James D. "Personal Wealth Transfers." *Q.J.E.* 95 (August 1980): 159–79.

Barlow, Robin; Brazer, Harvey E.; and Morgan, James N. *Economic Behavior of the Affluent.* Washington: Brookings, 1966.

Barro, Robert J. "Are Government Bonds Net Wealth?" *J.P.E.* 82, no. 6 (November/December 1974): 1095–1117.

Becker, Gary S. "A Theory of Social Interactions." *J.P.E.* 82, no. 6 (November/December 1974): 1063–94.

Becker, Gary S., and Lewis, H. Gregg. "On the Interaction between the Quantity and Quality of Children." *J.P.E.* 81, no. 2, pt. 2 (March/April 1973): S279–S288.

Becker, Gary S., and Tomes, Nigel. "Child Endowments and the Quantity and Quality of Children." *J.P.E.* 84, no. 4, pt. 2 (August 1976): S143–S162.

———. "An Equilibrium Theory of the Distribution of Income and Intergenerational Mobility." *J.P.E.* 87, no. 6 (December 1979): 1153–89.

Bevan, D., and Stiglitz, Joseph E. "The Wealth Distribution and Inheritance." Paper presented at American Economic Association meetings, Chicago, August 1978.

Blinder, Alan S. "A Model of Inherited Wealth." *Q.J.E.* 87 (November 1973): 608–26.

———. "Inequality and Mobility in the Distribution of Wealth." *Kyklos* 29, no. 4 (1976): 607–38. (*a*)

———. "Intergenerational Transfers and Life Cycle Consumption." *A.E.R. Papers and Proc.* 66 (May 1976): 87–93. (*b*)

Blomquist, N. S. "The Inheritance Function." *J. Public Econ.* 12 (August 1979): 41–60.

Brittain, John A. *Inheritance and the Inequality of Material Wealth.* Washington: Brookings, 1978.

Easterlin, Richard A. "Population Change and Farm Settlement in the Northern United States." *J. Econ. Hist.* 36 (March 1976): 45–75.

Edwards, Linda N., and Grossman, Michael. "An Economic Analysis of Children's Health and Intellectual Development." Working paper no. 180, Nat. Bur. Econ. Res., New York, May 1977.

Griliches, Zvi. "Sibling Models and Data in Economics: Beginnings of a Survey." *J.P.E.* 87, no. 5, pt. 2 (October 1979): S37–S64.

Harbury, Colin D., and Hitchens, D. M. W. N. *Inheritance and Wealth Inequality in Britain.* London: Allen & Unwin, 1979.

Heckman, James J. "A Life-Cycle Model of Earnings, Learning, and Consumption." *J.P.E.* 84, no. 4, pt. 2 (August 1976): S11–S44 (*a*)

———. "The Common Structure of Statistical Models of Truncation, Sample Selection and Limited Dependent Variables and a Simple Estimator for Such Models." *Ann. Econ. and Soc. Measurement* 5 (Fall 1976): 475–92. (*b*)

———. "Sample Selection Bias as a Specification Error." *Econometrica* 47 (January 1979): 153–61.

Ishikawa, Tsuneo. "Family Structures and Family Values in the Theory of Income Distribution." *J.P.E.* 83, no. 5 (October 1975): 987–1008.

Knight, Frank H. *Risk, Uncertainty and Profit.* 1921. Reprint. New York: Kelley & Millman, 1957.

Lebergott, Stanley. "Are the Rich Getting Richer? Trends in U.S. Wealth Concentration." *J. Econ. Hist.* 36 (March 1976): 147–62.

Leibowitz, Arleen. "Home Investments in Children." *J.P.E.* 82, no. 2, pt. 2 (March/April 1974): S111–S131.

Leibowitz, Arleen, and Friedman, Bernard. "Family Bequests and the Derived Demand for Health Inputs." *Econ. Inquiry* 17 (July 1979): 419–34.

Menchik, Paul L. "Inter-generational Transmission of Inequality: An Empirical Study of Wealth Mobility." *Economica* 46 (November 1979): 349–62.

———. "Primogeniture, Equal Sharing, and the U.S. Distribution of Wealth." *Q.J.E.* 94 (March 1980): 299–316.

Nerlove, Marc, and Press, S. James. "Univariate and Multivariate Log-linear and Logistic Models." Rand Report R-1306-EDA/NIH, Santa Monica, Calif., December 1973.

Parsons, Donald O. "Intergenerational Wealth Transfers and the Educational Decisions of Male Youth." *Q.J.E.* 89 (November 1975): 603–17. (*a*)

———. "On the Economics of Intergenerational Relations." Paper presented to the Workshop in the Applications of Economics, Univ. Chicago, 1975. (*b*)

Pryor, Frederic L. "Simulation of the Impact of Social and Economic Institutions on the Size Distribution of Income and Wealth." *A.E.R.* 63 (March 1973): 50–72.

Stamp, Josiah C. "Inheritance as an Economic Factor." *Econ. J.* 36 (September 1926): 339–74.

Stiglitz, Joseph E. "Notes on Estate Taxes, Redistribution, and the Concept of Balanced Growth Path Incidence." *J.P.E.* 86, no. 2, pt. 2 (April 1978): S137–S150.

Sussman, Marvin B.; Cates, Judith N.; and Smith, David T. *The Family and Inheritance.* New York: Russell Sage, 1970.

Tomes, Nigel. "A Model of Child Endowments, and the Quality and Quantity of Children." Ph.D. dissertation, Univ. Chicago, 1978.

———. "Uncertain Lifetimes, Imperfect Capital Markets and the Altruistic Motive for Bequests." Research Report no. 7930, Univ. Western Ontario, Dept. Econ., October 1979.

———. "A Model of Fertility and Children's Schooling." *Econ. Inquiry* (1981), in press. (*a*)

———. "On the Intergenerational Savings Function." *Oxford Econ. Papers* (1981), in press. (*b*)

Wedgwood, Josiah. "The Influence of Inheritance on the Distribution of Wealth: Some New Evidence." *Econ. J.* 38 (March 1929): 38–55.

Willis, Robert J., and Rosen, Sherwin. "Education and Self-Selection." *J.P.E.* 87, no. 5, pt. 2 (October 1979): S7–S36.

Intraindustry Specialization and the Gains from Trade

Paul R. Krugman

Massachusetts Institute of Technology

Several recent empirical studies of trade suggest that *interindustry* specialization and trade, which reflect the conventional forces of comparative advantage, are also accompanied by *intraindustry* specialization, which reflects scale economies and consumers' taste for a diversity of products. This paper develops a simple model which illustrates this argument. Two main results are developed. First, the nature of trade depends on how similar countries are in their factor endowments. As countries become more similar, the trade between them will increasingly become intraindustry in character. Second, the effects of opening trade depend on its type. If intraindustry trade is sufficiently dominant, the advantages of extending the market will outweigh the distributional effects, and the owners of scarce as well as of abundant factors will be better off.

Over the years, many empirical students of international trade have argued that trade among the industrial countries cannot adequately be explained by conventional theories of comparative advantage. One might summarize this empirical critique by pointing to three aspects of world trade which seem to contradict received theory. First, much of world trade is between countries with similar factor endowments. Second, a large part of trade is intraindustry in character—that is, it consists of two-way trade in similar products. Finally, much of the expansion of trade in the postwar period has taken place without

This research was supported by a grant from the National Science Foundation. An earlier version of this paper, "International Trade and Income Distribution. A Reconsideration," was presented at the NBER Summer Institute in International Studies, Cambridge, Massachusetts, July 1979.

[*Journal of Political Economy*, 1981, vol 89, no 5]
© 1981 by The University of Chicago. 0022-3808/81/8905-0005$01.50

sizable reallocation of resources or income-distribution effects. This last point is particularly noticeable in the cases of the EEC and the North American automobile pact.

The purpose of this paper is to formalize one possible explanation of these seeming paradoxes. The explanation is not a new one: It is essentially the same as that put forward by Balassa (1967), Grubel (1970), and Kravis (1971), among others. What this paper does is put the argument in terms of a formal model, a step which may be of some help in clarifying and disseminating ideas which have been "in the air" for some time.

Briefly, the argument of these empirical workers, a very clear exposition of which is given by Kravis (1971), runs as follows. The conventional forces of comparative advantage operate on *groups* of products ("industries") and thus give rise to *interindustry* specialization and trade. Economies of scale in production, however, lead each country to produce only a subset of the products within each group, so that there is also *intraindustry* specialization and trade. This provides a simple explanation of two of our empirical ostensible paradoxes. Countries with similar factor endowments will still trade because of scale economies, and their trade will be largely intraindustry in character. The third seeming paradox—the apparent painlessness of some trade liberalization—can also be resolved if we argue that income-distribution effects are outweighed by the gains from a larger market when countries are sufficiently similar.

While this is a simple and straightforward explanation, however, it is not so easy to formalize. Scale economies are crucial to the argument, and they are notoriously awkward to handle in general equilibrium models. In this paper I follow an earlier paper (Krugman 1979) and use the device of Chamberlinian monopolistic competition. As in the earlier paper, this proves to be a very convenient approach, yielding a simple and tractable model. The structure of this model and the determination of this model's equilibrium in a closed economy are set forth in Section I. Section II shows how the pattern of trade between two countries is determined in the model, developing the basic relationship between differences in factor endowments and the extent of intraindustry trade. Section III then examines the effects of trade on income distribution and shows how the extent of intraindustry trade determines whether scarce factors of production gain or lose from trade. Finally, Section IV summarizes the results and discusses some implications for theory and policy.

It must be emphasized that the model presented here is in no sense a general one. In addition to making strong assumptions about functional forms of cost and utility functions, I impose a great deal of symmetry on the model to simplify the analysis and give a natural

meaning to the concept of "similarity" in factor proportions. Thus the results of the analysis are at best suggestive. Nonetheless, they seem intuitively plausible and also seem to have something to do with actual experience.

I. The Model in a Closed Economy

Intraindustry trade depends on the existence of unexhausted economies of scale in production. The main problem in modeling this kind of trade is how to handle these scale economies, which must lead to a breakdown of perfect competition (unless they are wholly external to firms). In this paper, as in an earlier paper (Krugman 1979), I will use the device of Chamberlinian monopolistic competition, basing the model on recent work by Dixit and Stiglitz (1977). An "industry" will consist of a large number of firms, all producing somewhat differentiated products, all operating on the downward-sloping parts of their average cost curves. There will be two-way international trade within an industry because firms in different countries will produce different differentiated products. What prevents countries from producing a complete range of products domestically is the existence of fixed costs in production; thus scale economies are the basic cause of intraindustry trade.

We should note at the outset that the concept of an industry used in this paper is a somewhat special one. One might want to define an industry either as a group of products which are close substitutes on the supply side or as a group of products which are close substitutes on the demand side. In the model of this paper, I assume that there are two groups of products which fit both definitions. Products within each group are closer substitutes than products in different groups, while factors of production are assumed mobile among products within each group yet immobile between groups. This convenient coincidence of the two possible concepts of an industry may or may not be empirically reasonable; it is certainly not theoretically necessary and should be regarded as one among many special assumptions.

Another conceptual difficulty concerns the notion of a "product." In the formulation below, all products seem to look alike, since they enter symmetrically into cost and utility functions. This may seem to involve an illegitimate comparison of physical quantities of different goods. I show in the Appendix, however, that the formulation of many "identical" products can be interpreted as a restriction on the parameters of a model in which products really do differ.

Let us begin, then, with a two-industry model of a closed economy. Each industry consists of a large number of products, all of which enter symmetrically into demand, with the two industries—industry 1

and industry 2—themselves playing symmetric roles. All individuals will have the convenient utility function.

$$U = \ln\left(\sum_{i=1}^{N_1} c_{1,i}^{\theta}\right)^{1/\theta} + \ln\left(\sum_{j=1}^{N_2} c_{2,j}^{\theta}\right)^{1/\theta}, \qquad (1)$$
$$0 < \theta < 1,$$

where $c_{1,i}$ is consumption of the ith product of industry 1; $c_{2,j}$ is consumption of the jth product of industry 2; and N_1 and N_2 are the (large) numbers of potential products in each industry. Not all potential products will necessarily be produced, and we will in fact assume that the actual numbers of products produced—n_1 and n_2—while large, fall short of N_1 and N_2.

The utility function (1) has several useful properties. First, it ensures that half of income will always be spent on industry 1's products. Second, if the number of products in each industry is large, it implies that every producer faces a demand curve with elasticity $1/(1 - \theta)$. Finally, (1) will allow us to represent the gains and losses from trade in a particularly simple way.

On the demand side, then, an industry is assumed to consist of a number of products which are *imperfect* substitutes for one another. On the supply side, however, they will be assumed to be *perfect* substitutes. There will be only two factors of production, type 1 labor and type 2 labor, each of which is wholly specific to an industry but nonspecific among products within an industry. Thus, type 1 labor will be used only in industry 1, type 2 only in industry 2. Within each industry, the labor required to produce a particular product will consist of a fixed setup cost and a constant variable cost;

$$\begin{aligned} l_{1,i} &= \alpha + \beta x_{1,i}, \qquad i = 1, \ldots, n_1, \\ l_{2,j} &= \alpha + \beta x_{2,j}, \qquad j = 1, \ldots, n_2, \end{aligned} \qquad (2)$$

where $l_{1,i}$ is labor used in producing the ith product of industry, 1; $x_{1,i}$ is the output of that product; and so on. To go from these required labor inputs to nominal costs, we must multiply by the wage rates of the two types of labor, w_1 and w_2.

To close the model, we begin by noting that output of each product, x, is the sum of individual consumptions of the product. At the same time, total employment in each industry is the sum of employment in producing all the individual products. Assuming full employment, we have

$$\left.\begin{aligned}\sum_{i=1}^{n_1} l_{1,i} &= L_1 = 2 - z\\ \sum_{j=1}^{n_2} l_{2,j} &= L_2 = z\end{aligned}\right\} 0 < z < 1. \tag{3}$$

Thus the total labor force is set equal to 2, with the parameter z measuring factor proportions. As we will see below, z will assume crucial significance in determining the importance of intraindustry trade and the effect of trade on income distribution.

We are now prepared to examine the determination of equilibrium in this model. This involves determining how many products are actually produced in each industry, the output of each product, the prices of products, and the relative wages of the two kinds of labor. We should note at the outset that it is indeterminate *which* products are produced—but it is also unimportant.

Our first step is to determine the pricing policy of firms. We assume that producers can always costlessly differentiate their products. This means that each product will be produced by only one firm. If there are many products the elasticity of demand for each product will, as already noted, be $1/(1 - \theta)$. (This is proved in the Appendix.) Thus, each firm will face a demand curve of constant elasticity. We then have the familiar result that the profit-maximizing price will be marginal cost plus a constant percentage markup:

$$\begin{aligned}p_1 &= \theta^{-1}\beta w_1,\\ p_2 &= \theta^{-1}\beta w_2,\end{aligned} \tag{4}$$

where p_1 and p_2 are the prices of any products in industry 1 and 2, respectively, which are actually produced.

Given the pricing policy of firms, actual profits depend on sales:

$$\begin{aligned}\pi_1 &= p_1x_1 - (\alpha + \beta x_1)w_1,\\ \pi_2 &= p_2x_2 - (\alpha + \beta x_2)w_2,\end{aligned} \tag{5}$$

where x_1 and x_2 are sales of representative firms in the two industries.

But in this model there will be free entry of firms, driving each industry to Chamberlin's "tangency solution" where profits are zero. Thus we can use the condition of zero profits in equilibrium to determine the equilibrium size and number of firms. Setting $\pi_1 = \pi_2 = 0$ and using (4) and (5), we have

$$x_1 = x_2 = \frac{\alpha}{\beta} \cdot \frac{\theta}{1 - \theta} \tag{6}$$

for the size of firms. The number of firms can then be determined from the full-employment condition:

$$n_1 = (2 - z)/(\alpha + \beta x_1),$$
$$n_2 = z/(\alpha + \beta x_2). \tag{7}$$

The final step in determining equilibrium is to determine relative wages. This can be done very simply by noting that the industries receive equal shares of expenditure and that, since profits are zero in equilibrium, these receipts go entirely to the wages of the industry-specific labor forces. So $w_1L_1 = w_2L_2$, implying

$$w_1/w_2 = z/(2 - z). \tag{8}$$

We now have a completely worked out equilibrium for a two-sector, monopolistically competitive economy. It is indeterminate which of the range of potential products within each industry are actually produced, but since all products appear symmetrically, this is of no welfare significance. The character of the economy is determined by the two parameters z and θ. The value of z determines relative wages: If z is low, type 2 labor will receive much higher wages than type 1 labor. The value of θ measures the degree of substitutability among products within an industry. The lower is θ, the more differentiated are products, and the more important are unexploited scale economies. From (4) we have $\theta = \beta w_1/p_1 = \beta w_2/p_2$. But βw_1 and βw_2 are the marginal costs of production, while in equilibrium price equals average cost. Thus θ is the ratio of marginal to average cost (which is also the elasticity of cost with respect to output).

II. Factor Proportions and the Pattern of Trade

In the last section we saw how equilibrium can be determined in a simple closed-economy model with scale economies and differentiated products. We can now examine what happens when two such economies trade. What we are principally concerned with is the proposition, advanced in the introduction, that countries with similar factor endowments will engage in intraindustry trade, while countries with very different endowments will engage in Heckscher-Ohlin trade.

As a first step we need a working measure of the extent of intraindustry trade. The empirical literature on intraindustry trade (e.g., Hufbauer and Chilas 1974; Grubel and Lloyd 1975) generally concentrates on an index of trade overlap, that is,

$$I = 1 - \left(\sum_k |X_k - M_k|\right) \Big/ \left[\sum_k (X_k + M_k)\right], \tag{9}$$

where X_k is a country's exports in industry k and M_k is imports in that industry. This index has the property that, if trade is balanced industry by industry, it equals one, while if there is complete international specialization so that every industry is either an export or import industry, it equals zero. As we will see, this index fits in quite well with the model of this paper.

The other concept we need to make operational is that of similarity in factor endowments. In general, this is not well defined. What I will do in this paper, however, is consider a special case in which the concept does have a natural meaning without trying to arrive at a general definition.

Let us suppose, then, that there are two countries, the home country and the foreign country. The home country will be just as described in Section I. The foreign country will be identical except for one thing: The relative sizes of the two industries' labor forces will be reversed. That is, the foreign country will be a mirror image of the home country. If we use a star on a variable to indicate that it refers to the foreign country, we have

$$\begin{aligned} L_1 &= 2 - z \qquad & L_2 &= z \\ L_1^* &= z \qquad & L_2^* &= 2 - z. \end{aligned} \tag{10}$$

Obviously, given this pattern of endowments, we can regard z as an index of similarity in factor proportions. If $z = 1$, the countries have identical endowments. As z gets smaller, the factor proportions become increasingly different.

The mirror-image assumption can be given a geometric interpretation. In figure 1, an Edgeworth box is used to represent the international distribution of productive resources. The origin O is used to measure home country endowments, O^* to measure foreign endowments. The two diagonals of the box can then be given economic interpretations: OO^* is a line along which factor proportions are equal in the two countries, while the other diagonal is a line along which the countries are of equal economic size. The mirror-image assumption is saying that the endowment point E lies on this diagonal. The parameter z then determines the position of E; as z goes from 0 to 1, E moves from the corner to the center of the box.

Suppose, now, that these countries are able to trade at zero transportation cost. As before, we can determine pricing behavior, the size and number of firms, and relative wages. In addition, we can determine the volume and pattern of trade.

The first point to note is that the elasticity of demand for any particular product is still $1/(1 - \theta)$. This gives us price equations

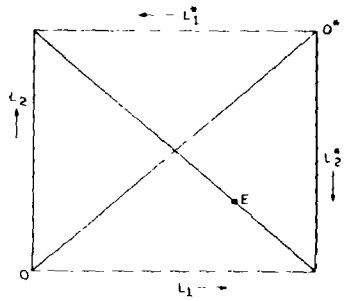


FIG. 1

exactly the same as before:

$$\begin{aligned} p_1 &= \theta^{-1}\beta w_1, \\ p_2 &= \theta^{-1}\beta w_2, \\ p_1^* &= \theta^{-1}\beta w_1^*, \\ p_2^* &= \theta^{-1}\beta w_2^*. \end{aligned} \tag{11}$$

Now, however, the symmetry of the setup insures that all wages will be equal, both across industries and internationally:

$$w_1 = w_1^* = w_2 = w_2^*. \tag{12}$$

The zero-profit condition will determine the equilibrium size of firm, x, which will be the same for both industries in both countries:

$$x = \alpha\theta/\beta(1 - \theta). \tag{13}$$

Finally, full employment determines the number of firms in each industry in each country:

$$\begin{aligned} n_1 &= n_2^* = (2 - z)/(\alpha + \beta x), \\ n_2 &= n_1^* = z/(\alpha + \beta x). \end{aligned} \tag{14}$$

What these results show is that trade will lead to factor price equalization while leaving the pattern of production unchanged. Our remaining task is to determine the volume and pattern of trade. We can do this by noting two points. First, everyone will devote equal shares of expenditure to the two industries. Second, everyone will spend an equal amount on each of the products within an industry. This means that the share of all individuals' income falling on, say, industry 1 products produced in the foreign country is $\frac{1}{2} \cdot [n_1^*/(n_1 + n_1^*)]$—that is, the industry share in expenditure times that country's share of the industry. But the number of products is proportional to the labor force. Thus, if we let Y be the home country's income

(equal to the foreign country's), X_1 be exports of industry 1 products, X_2 be exports of industry 2 products, M_1 be imports of industry 1 products, and M_2 be imports of industry 2 products, we have

$$\begin{aligned} X_1 &= \tfrac{1}{2}Y \cdot [(2 - z)/2], \\ X_2 &= \tfrac{1}{2}Y \cdot (z/2), \\ M_1 &= \tfrac{1}{2}Y \cdot (z/2), \\ M_2 &= \tfrac{1}{2}Y \cdot [(2 - z)/2]. \end{aligned} \tag{15}$$

Now, the relations (15) have two important implications. First, consider the volume of trade. Total home country exports are $X_1 + X_2 = \tfrac{1}{2}Y$. Thus the ratio of trade to income is independent of z, the index of similarity in factor proportions. This can be regarded as an answer to the first ostensible empirical paradox mentioned in the introduction—the large volume of trade among similar countries. In this model, similar countries will trade just as much as dissimilar countries.

The second seeming empirical paradox was the prevalence, in trade among similar countries, of two-way trade in similar products. If we substitute (15) into our expression for intraindustry trade (9), we get a simple, striking result:

$$I = z. \tag{16}$$

The index of intraindustry trade equals the index of similarity in factor proportions.

These results may appear to depend crucially on the assumptions of this model, but in qualitative terms they can survive a good deal of generalization. The persistence of trade between countries with similar factor endowments will occur in almost any model with economies of scale. The relationship between similarity of countries and the extent of intraindustry trade can be shown to hold, for an appropriate definition of similarity, in a much more general model and has also been noted in a quite different context by Ethier (1979). Insofar as these insights are concerned, the virtue of this model is not in the difference of its conclusions but in the clarity with which they emerge.

Where the special assumptions of this model become particularly useful, however, is in attempting to deal with the welfare consequences of trade. These consequences are considered in the next section.

III. Gains and Losses from Trade

In this section we must again begin by delineating a concept which I have been using loosely. This is the idea of the "seriousness" of

distribution problems. What we need is a clear way of formulating the notion that distribution problems from opening trade will not be serious, if countries are sufficiently similar in factor proportions that the trade which results is primarily intraindustry trade.

The criterion I will use to define nonserious distribution problems is the following: Distribution problems arising from trade will be held not to be serious if *both factors gain from trade*. This, of course, begs some questions, since there may be difficulties in getting groups to accept a relative decline in income even if they are absolutely better off. But this criterion is fairly reasonable and turns out to give suggestive results.

To find out whether factors gain from trade, we need to know how utility depends on the variables of the model. Suppose an individual receives a wage w and has the utility function (1). He will then spend $w/2$ on the products of each industry and divide his expenditure equally among the products within an industry. Thus his utility will depend on his wage, the prices of representative products in each industry, and the number of products available:

$$\begin{aligned} U &= \ln [n_1(w/2n_1p_1)^\theta]^{1/\theta} + \ln [n_2(w/2n_2p_2)^\theta]^{1/\theta} \\ &= 2 \ln 2 + \ln w/p_1 + \ln w/p_2 + \frac{1-\theta}{\theta} \ln n_1 + \frac{1-\theta}{\theta} \ln n_2. \end{aligned} \tag{17}$$

The function (17) has the convenient property that all the effects enter additively. Utility depends on real wages in terms of representative products and on diversity.

To analyze the effects of trade on welfare, it is useful to introduce some more notation:

U_1, U_2 = utility of workers in industries 1 and 2;

w_{11}, w_{12} = real wage of industry 1 workers in terms of products of industries 1 and 2;

w_{21}, w_{22} = real wage of industry 2 workers in terms of products of industries 1 and 2.

Then we can substitute into (17) to get (suppressing the constant term):

$$\begin{aligned} U_1 &= \ln w_{11} + \ln w_{12} + \frac{1-\theta}{\theta} \ln n_1 + \frac{1-\theta}{\theta} \ln n_2, \\ U_2 &= \ln w_{21} + \ln w_{22} + \frac{1-\theta}{\theta} \ln n_1 + \frac{1-\theta}{\theta} \ln n_2. \end{aligned} \tag{18}$$

We are now in a position to measure the welfare effects of trade. Suppose we start from a position of autarky, as in Section I, then move to free trade, as in Section II. There will then be two kinds of

effects. First, there will be a distribution effect as factor prices are equalized. As one can easily verify, labor's real wage remains the same in terms of the products of its own industry while rising or falling in terms of the other industry's products, depending on whether the factor is abundant or scarce. Thus, in the home country this effect benefits labor in industry 1 and hurts labor in industry 2.

The second effect comes from the increase in the size of the market, which makes a greater variety of products available. This works to everyone's benefit.

Since both effects work in its favor, the abundant factor must be made better off. This leaves us with the problem of determining the change in utility of the scarce factor—industry 2 labor in the home country and the symmetrically placed industry 1 labor in the foreign country.

Let a prime on a variable indicate its free-trade value while unmarked variables refer to autarky. Then, as we move from the autarky solution in Section I to the free-trade solution in Section II, the change in U_2 is

$$\begin{aligned} U_2' - U_2 &= \ln w_{21}'/w_{21} + \frac{1-\theta}{\theta}\ln n_1'/n_1 + \frac{1-\theta}{\theta}\ln n_2'/n_2 \\ &= \ln z/(2-z) + \frac{1-\theta}{\theta}\ln 2/(2-z) + \frac{1-\theta}{\theta}\ln 2/z, \end{aligned} \tag{19}$$

where the first term is negative and represents the distribution loss; the remaining terms are positive and represent the gains from being part of a larger market. The question is under what conditions these terms will outweigh the first terms.

By collecting terms, we can rewrite (19) as

$$U_2' - U_2 = \frac{2\theta - 1}{\theta}\ln z - \frac{1}{\theta}\ln 2 - z + \frac{2-2\theta}{\theta}\ln 2. \tag{20}$$

This gives us one immediate result: If $\theta \leq 0.5$, the scarce factor necessarily gains from trade, since the first term will be positive and the third term will outweigh the second. Recall that θ is a measure of the substitutability of products within an industry. What this result then says is that *if products are sufficiently differentiated, both factors gain from trade.*

If $\theta > 0.5$, whether both factors gain depends on the extent to which trade is intraindustry in character, which in turn depends on *how similar the countries are in factor proportions.* When $\theta > 0.5$, the function (20) has three properties: (i) as z approaches 1, $U_2' - U_2$ goes to $[(2-2\theta)/\theta]\ln 2 > 0$; (ii) as z goes to zero, $U_2' - U_2$ goes to minus infinity; and (iii) $U_2' - U_2$ is strictly increasing in z. Thus, if we were to graph (20), it would look like figure 2. There is a critical value of z, ξ,

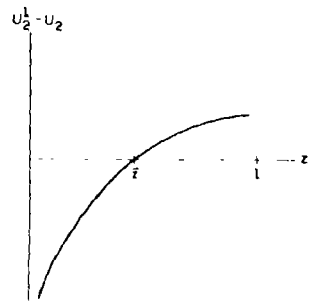


FIG. 2

for which $U_2' - U_2 = 0$. If $z > \bar{z}$, both factors gain; if $z < \bar{z}$, the scarce factor loses. But z is our measure of similarity in factor proportions. Thus what we have shown is that *if countries have sufficiently similar factor endowments, both factors gain from trade.*

What is particularly nice about this result is that we have already seen that there is a one-for-one relationship between similarity of factor endowments and intraindustry trade. So this result can be taken as a vindication of the arguments of such authors as Kravis (1971) and Hufbauer and Chilas (1974) that intraindustry trade poses fewer adjustment problems than interindustry trade.

We should note, however, that the critical value of interindustry trade depends on the substitutability of products. The function (20) is decreasing in θ: $\partial(U_2' - U_2)/\partial\theta = \theta^{-2} \ln z(2 - z) < 0$. So an increase in θ will shift the function down. This will increase $\bar{z}$. *The less differentiated are products, the more similar countries must be if both factors are to gain from trade.* In the limit, as θ goes to 1, so does $\bar{z}$.

The results of this section are summarized in figure 3. On the axes are the two parameters θ and z, both capable of taking on values

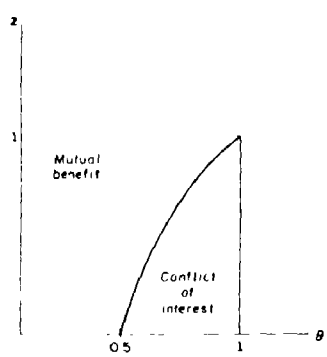


FIG. 3

between zero and one. What we have shown is that the qualitative effects of trade depend on where we are in the unit square. In the southeastern part of the square—labeled "conflict of interest"—either scale economies are unimportant or countries are very different in factor endowments, and scarce factors lose from trade. In the other region—"mutual benefit"—the gains from intraindustry specialization outweigh the conventional distributional effects, and everyone gains from trade.

IV. Summary and Conclusions

This paper began with three "paradoxes" about international trade. Since they do not seem paradoxical in the light of this model, perhaps we should state them as "stylized facts": (i) Much of world trade is between countries with similar factor endowments. (ii) The trade between similar countries is largely intraindustry in character; that is, it consists of two-way trade in similar products. (iii) The growth of intraindustry trade has not posed serious income-distribution problems.

This paper offers a simple model which formalizes one possible explanation of these stylized facts. According to this view, the variety of products produced in any one country is limited by the existence of scale economies in production. Thus similar countries have an incentive to trade; their trade will typically be in products produced with similar factor proportions; and this trade will not involve the income-distribution effects characteristic of more conventional trade.

In addition to helping make sense of some puzzling empirical results, this paper is, I hope, of some interest from the standpoint of pure theory. The model dispenses with the two most fundamental assumptions of standard trade theory: perfect competition and constant returns to scale. Instead, I have dealt in this paper with a world in which economies of scale are pervasive, and all firms have monopoly power. While the model depends on extremely restrictive assumptions, it does show that it is possible for trade theory to make at least some progress into this virtually unexplored territory.

Appendix

1. *The Concept of a Product*

In the formulation in Section I, an industry was assumed to consist of many products with the "same" cost function and entering in the "same" way into utility. This may seem to involve a comparison of apples and oranges. However, it can be justified as a restriction on the parameters of a more general model.

Consider the utility and cost functions for a one-industry model (the generalization to two industries is obvious):

$$U = \left[\sum_{i=1}^{N} (\delta_i C_i)^{\theta}\right]^{1/\theta}, \; 0 < \theta < 1, \tag{A1}$$

$$l_i = \alpha_i + \beta_i X_i, \quad i = 1, \ldots, n. \tag{A2}$$

Here we allow goods to enter with different weights into utility and to have different cost functions; thus no assumption is made about comparability of units. Given certain restrictions on parameters, however, it is possible to choose units so that a formulation where all products appear identical is valid. Let us suppose first that $\alpha_i = \alpha$ for all i. The measurement of this cost is independent of the choice of units, so this is a meaningful assumption. Let us also assume $\beta_i/\delta_i = \beta$ for all i. This again does not depend on units of measurement; measuring product 27 in batches of 10 instead of individual units will increase both β_{27} and δ_{27} by a factor of 10 and leave the ratio unchanged.

If the assumptions about parameters are granted—and they are special assumptions, not general properties—we can justify the model in the text by a choice of units. Let $\hat{C}_i = \delta_i C_i$ for all i. Then the utility and cost functions become

$$U = \left(\sum_{i=1}^{N} \hat{C}_i^{\theta}\right)^{1/\theta} \tag{A3}$$

$$l_i = \alpha + \beta \hat{X}_i, \quad i = 1, \ldots, n \tag{A4}$$

II. *Elasticity of Demand for Individual Products*

The analysis in Section I depends on the result that the elasticity of demand for any particular product is $1/(1-\theta)$. This Appendix gives a demonstration of this.

Consider an individual maximizing his utility function (1) subject to a budget constraint. The first-order conditions from that maximization will have the form

$$p_{1,i} = \frac{c_{1i}^{(1-\theta)}}{\lambda \sum_k c_{1k}^{\theta}}, \quad i = 1, \ldots, n_1$$

$$p_{2,j} = \frac{c_{2j}^{(1-\theta)}}{\lambda \sum_m c_{2m}^{\theta}}, \quad j = 1, \ldots, n_2,$$

where λ is the shadow price on the budget constraint, that is, the marginal utility of income.

If there are many products, however, the firm producing a particular product can take the denominators of these expressions as given. Thus each individual's demand for a particular product, and therefore also market demand, will have elasticity $1/(1-\theta)$.

References

Balassa, Bela. *Trade Liberalization among Industrial Countries: Objectives and Alternatives*. New York: McGraw-Hill, 1967.

Dixit, Avinash K., and Stiglitz, Joseph E. "Monopolistic Competition and Optimum Product Diversity." *A.E.R.* 67 (June 1977): 297–308.

Ethier, Wilfred. "Internationally Decreasing Costs and World Trade." *J. Internat. Econ.* 9 (February 1979): 1–24.

Grubel, Herbert G. "The Theory of Intra-Industry Trade." In *Studies in International Economics*, edited by I. A. McDougall and Richard H. Snape. Amsterdam: North-Holland, 1970.

Grubel, Herbert G., and Lloyd, Peter J. *Intra-Industry Trade. The Theory and Measurement of International Trade in Differentiated Products.* New York: Wiley, 1975.

Hufbauer, Gary C., and Chilas, John G. "Specialization by Industrial Countries: Extent and Consequences." In *The International Division of Labour: Problems and Perspectives*, edited by Herbert Giersch. Tubingen: Mohr, 1974.

Kravis, Irving B. "The Current Case for Import Limitations." In *United States Economic Policy in an Interdependent World*, Commission on International Trade and Investment Policy. Washington: Government Printing Office, 1971.

Krugman, Paul R. "Increasing Returns, Monopolistic Competition, and International Trade." *J. Internat. Econ.* 9 (November 1979): 469–79.

The Adjustment of Consumption to Changing Expectations about Future Income

Marjorie A. Flavin

University of Virginia

The paper analyzes the role of current income in providing new information about future income and thus signaling changes in permanent income. Using time-series analysis to quantify the revision in permanent income induced by an innovation in the current income process, a structural econometric model of consumption is developed. The rejection of the joint rational expectations–permanent income hypothesis is both statistically and quantitatively significant. The paper also shows that the test of the rational expectations–permanent income hypothesis proposed by Hall is based on the reduced form of this structural model and reconciles Sargent's consumption paper with Hall's.

Because the path of future income is uncertain, the individual's permanent income and, therefore, his consumption plans will be revised period by period as new information about future income becomes available. Under the assumption that individuals hold rational expectations of future income, the stochastic properties of these period-by-period revisions in permanent income can be exploited in order to formulate testable restrictions on the joint time-series behavior of consumption and other observable variables.

This paper first reviews the econometric tests of the permanent income–rational expectations hypothesis proposed by Hall (1978) and Sargent (1978), then develops and executes an alternative approach to testing the hypothesis. Section I formally defines permanent in-

This paper is part of my doctoral dissertation at MIT. I am grateful to Robert Hall for guidance, and to Andrew Abel, Stanley Fischer, Robert Flood, Peter Garber, Robert Lucas, Bennett McCallum, Jeffrey Sachs, Kenneth Singleton, Richard Startz, and several referees for helpful suggestions.

[*Journal of Political Economy*, 1981, vol. 89, no. 5]
© 1981 by The University of Chicago. 0022-3808/81/8905-0003$01.50

come and derives the basic implication of the permanent income–rational expectations hypothesis: If consumption is proportional to permanent income in each period, and permanent income, in turn, represents the best estimate, given currently available information, of the individual's lifetime resources, then current consumption should diverge from consumption in the previous period by the amount of the contemporaneous revision in permanent income.

The published papers of Hall (1978) and Sargent (1978) both analyze this predicted relationship between consumption, lagged consumption, and the revision in permanent income, conditional on lagged information sets.[1] The two papers reach conflicting empirical conclusions, Hall confirming and Sargent decisively rejecting the permanent-income hypothesis. Section II argues that an incorrect definition of permanent income used in Sargent's formulation of the restrictions is the source of the discrepancy and shows that Sargent's analysis, once the definition of permanent income is corrected, leads to the same parameter restrictions tested by Hall.

The restrictions tested by Hall and Sargent apply to regressions in which all of the right-hand-side variables are lagged at least 1 period with respect to the dependent variable, consumption. By analyzing the expectation conditional on lagged information sets, their approach makes use of the property that the revision in permanent income should be uncorrelated with lagged variables. The Hall-Sargent tests can be thought of as tests based on a reduced-form consumption equation, since a reduced form expresses the expectation of an endogenous variable, conditional on predetermined variables.

This paper takes an alternative approach to testing the permanent income–rational expectations hypothesis by explicitly specifying a structural model of consumption. The model developed in Section III analyzes the role of current income in providing new information about future income and thus signaling changes in permanent income. The model is intended to underscore the point that the observed sensitivity of consumption to current income should not be interpreted simply as the behavioral marginal propensity to consume

[1] Bilson and Glassman (1979) consider some extensions on the model which I develop in this paper, Bilson (1979) reports comparative estimates of the model applied to data for the United States, United Kingdom, and Germany. Barro (1978) has also addressed the question of whether certain variables affect consumption only through their role in predicting future income. Recent empirical work on consumption by Blinder (1981) incorporates explicit treatment of the expectations of future income in the permanent-income model. Blinder's analysis concentrates on measuring the response of consumer expenditure to temporary changes in taxes rather than testing the permanent-income hypothesis itself. Hayashi (1979) also tests the permanent-income hypothesis and finds that consumption is more sensitive to current income than predicted by the hypothesis.

out of current income. Because income is a fairly highly serially correlated process, the fluctuations in current income will be correlated with fluctuations in permanent income. Formally, the model uses an autoregressive–moving average (ARMA) representation of the income time series in order to quantify the magnitude of the revision in permanent income implied by the contemporaneous observation of current income. The forecast error, or innovation, in the income time-series process represents the new information contained in the current observation on income. Based on the realization of the current innovation in the income process, individuals revise their expectations about future income. The magnitude of the revisions in expectations of future income in response to the realization of the current innovation depends on the parameters of the ARMA representation of the income process. Therefore, the revision in permanent income in a given period will be proportional to the innovation in current income, with the factor of proportionality determined by the parameters of the income time-series process. Following this approach, one can specify a structural equation relating the change in consumption to the contemporaneous revision in permanent income (modeled using the income innovation) and the change in current income. The coefficient of the change in income is a measure of the behavioral marginal propensity to consume out of current income, since the role of current income in signaling changes in permanent income has been explicitly modeled. The permanent income–rational expectations hypothesis can then be tested by testing whether *the marginal propensity to consume out of current income is significantly* different from zero. Having developed a test of the permanent income–rational expectations hypothesis in the context of a simple structural econometric system, it is easy to show that the Hall-Sargent test is the equivalent test based on the reduced form of the system. If the structural system is just identified, the test based on the structural system and the Hall-Sargent test based on the reduced form yield numerically identical values of the test statistic for the permanent income–rational expectations hypothesis. While the structural approach to testing the hypothesis is not statistically more powerful than the reduced-form approach, it is considerably more illuminating since it permits the recovery of the structural parameters of the system.

The empirical results indicate that the observed sensitivity of consumption to current income is greater than is warranted by the permanent income–life cycle hypothesis, even when the role of current income in signaling changes in permanent income is taken into account. The restrictions implied by the permanent income–rational expectations hypothesis can be rejected statistically at very low

confidence levels. Further, the estimates of the marginal propensity to consume out of current income are quite large, indicating that the failure of the permanent-income hypothesis is quantitatively, as well as statistically, significant.

I

In its simplest form, the permanent-income hypothesis holds that individuals ought to gear their consumption to their lifetime resources rather than to their measured income in some arbitrary time period. To start with a particularly simple case, consider the consumption behavior of an infinitely lived individual in a world in which the return to real wealth is certain and constant over time. The total resources available to the individual for current and future consumption consist of current real wealth, w_t, plus current and future labor income, x_{t+s}, $s = 0, 1, 2, \ldots, \infty$. Labor income is assumed to be exogenously determined. The individual's lifetime resources can be represented in either stock or flow form—the stock form being net worth, or the sum of real wealth and the present discounted value of current and future labor income, and the flow form being permanent income, or the annuity value of net worth.

Since the path of future income is not known with certainty, the individual must make his consumption plans on the basis of some set of expectations about future income. Given the expectations about future income held in period t, the individual's permanent income can be expressed as

$$y_t^p = r\left[w_t + \sum_{s=0}^{\infty}\left(\frac{1}{1+r}\right)^{s+1} E_t x_{t+s}\right] \tag{1}$$

In this discrete time formulation, w_t represents real wealth at the beginning of period t, x_t is assumed to be paid at the end of the period; and the real rate of return, r, is assumed constant.

Permanent income can be thought of as the constant resource flow which, conditional on expectations in period t, can be sustained for the remainder of the individual's time horizon. Thus, permanent income ought to have the property

$$E_t y_{t+1}^p = y_t^p. \tag{2}$$

To verify that permanent income, as specified in equation (1), has this property, note that real wealth will evolve according to

$$w_{t+1} = (1 + r)w_t + x_t - y_t^p \tag{3}$$

if the individual's consumption is identically equal to his permanent

discounted future income is not the appropriate definition of permanent income—is not affected by relaxing the restriction that $B = 1$. In order to accommodate Sargent's more general consumption function, the analysis of Section I will be generalized to allow for an unrestricted MPC.

Based on the definition of permanent income as the annuity value of net worth, equation (1), and the consumption function with unrestricted MPC, c_{t+1} will be related to c_t as follows:

$$c_{t+1} = [1 + r(1 - B)]c_t + rB \sum_{s=0}^{\infty} \left(\frac{1}{1 + r}\right)^{s+1} (E_{t+1} - E_t)x_{t+s+1} \quad (15)$$

$$- (1 + r)u_t + u_{t+1}.$$

If $B \neq 1$, consumption will be a random walk with drift instead of a simple random walk. Taking the expectation in period $t - 1$ gives:

$$E_{t-1}c_{t+1} = [1 + r(1 - B)]E_{t-1}c_t - (1 + r)E_{t-1}u_t + E_{t-1}u_{t+1}. \quad (16)$$

In comparing equation (16) with Sargent's equation for $E_{t-1}c_{t+1}$, (14), remember that $1/\alpha = 1 + r$.

The presence of the term $B(1 - \alpha)E_{t-1}y_t$ in Sargent's equation for $E_{t-1}c_{t+1}$ is an artifact of his definition of permanent income as average future income. Sargent's income series, y_t, is disposable income and therefore includes future nonlabor income as well as future labor income. However, the present discounted value of the future path of nonlabor income is not equal to current nonhuman wealth. Consider the definition of permanent income as the annuity value of net worth:

$$y_t^p = r\left[w_t + \sum_{s=0}^{\infty} \left(\frac{1}{1 + r}\right)^{s+1} E_t x_{t+s}\right], \quad (17)$$

where x_{t+s} is future labor earnings. The nonhuman wealth term, w_t, could be expressed as the present value of a future stream of earnings rather than as a current stock. Nonhuman wealth w_t, if maintained at that level forever, would yield a stream of nonlabor income of rw_t each period. Thus,

$$w_t = \sum_{s=0}^{\infty} \left(\frac{1}{1 + r}\right)^{s+1} rw_t. \quad (18)$$

Note, however, that the nonlabor income in the present value term is rw_t, the future nonlabor income the individual would receive each period *if* nonhuman wealth were maintained at w_t. Expanding equation (18) to isolate actual nonlabor income in each future period, rw_{t+s}, gives

$$w_t = \sum_{s=0}^{\infty} \left(\frac{1}{1+r}\right)^{s+1} r(w_t - w_{t+s}) + \sum_{s=0}^{\infty} \left(\frac{1}{1+r}\right)^{s+1} r w_{t+s}. \tag{19}$$

Equation (19) distinguishes between two conceptually different future revenue flows associated with w_t. The second term is simply the present value of future nonlabor earnings, which of course depends on the path of nonhuman wealth over time. The first term reflects receipts from the sale of the wealth's principal and payments to increase the principal. By defining permanent income as discounted future income, Sargent includes the second term but omits the first. Since the individual will systematically accumulate and decumulate nonhuman wealth over time in order to smooth his planned consumption path, the first term in equation (19) will not be identically zero or even constant over time. It is the omission of the term $\Sigma_{s=0}^{\infty} [1/(1+r)]^{s+1} r(w_t - w_{t+s})$ from his definition of permanent income which is responsible for the discrepancy between Sargent's version of the restrictions implied by the permanent-income hypothesis (eq. [14]) and the corrected version of the restrictions (eq. [16]).

Sargent's procedure is to impose the restrictions across the parameters of the vector autoregression of (c_t, y_t) implied by equation (14). The unrestricted autoregression of (c_t, y_t) is:

$$E_{t-1}c_t = c_1 y_{t-1} + c_2 y_{t-2} + \ldots + c_n y_{t-n} + d_1 c_{t-1} + d_2 c_{t-2} + \ldots + d_n c_{t-n} \tag{20}$$

and

$$E_{t-1}y_t = a_1 y_{t-1} + a_2 y_{t-2} + \ldots + a_n y_{t-n} + b_1 c_{t-1} + b_2 c_{t-2} + \ldots + b_n c_{t-n} \tag{21}$$

Since $E_{t-1}c_{t+1} = c_1 E_{t-1}y_t + c_2 y_{t-1} + \ldots + c_n y_{t-n+1} + d_1 E_{t-1}c_t + d_2 c_{t-1} + \ldots + d_n c_{t-n+1}$, substituting out for $E_{t-1}y_t$ and $E_{t-1}c_t$ yields an equation for $E_{t-1}c_{t+1}$ in terms of y_{t-i} and c_{t-i}, $i = 1, n$, and the a_i, b_i, c_i, d_i parameters. Substituting these autoregressive representations for $E_{t-1}c_t$, $E_{t-1}y_t$, and $E_{t-1}c_{t+1}$ into equation (14) gives a set of restrictions across the a_i, b_i, c_i, d_i parameters. These restrictions are tested by estimating equations (20) and (21) with and without the cross-equation restrictions imposed and computing a likelihood ratio statistic.

Following Sargent's procedure, but substituting the autoregressive representations of $E_{t-1}c_t$ and $E_{t-1}c_{t+1}$ into equation (16), the "cor-

rected" version of equation (14), provides the following set of restrictions on the parameters of the (c_t, y_t) vector autoregression:

$$\begin{aligned}
&c_1a_1 + c_2 + d_1c_1 = [1 + r(1 - B)]c_1 \\
&\cdots \\
&c_1a_{n-1} + c_n + d_1c_{n-1} = [1 + r(1 - B)]c_{n-1} \\
&c_1a_n + d_1c_n = [1 + r(1 - B)]c_n \\
&c_1b_1 + d_1^2 + d_2 = [1 + r(1 - B)]d_1 \\
&\cdots \\
&c_1b_{n-1} + d_1d_{n-1} + d_n = [1 + r(1 - B)]d_{n-1} \\
&c_1b_n + d_1d_n = [1 + r(1 - B)]d_n.
\end{aligned} \tag{22}$$

This modified set of restrictions could be tested, following Sargent's procedure, with a likelihood ratio test. However, rather than test the restrictions in their present form, I will argue that the time-series implications of the rational-expectations and permanent-income hypotheses can be stated and tested in a simpler form.

The restrictions in equation (22) are based on the projection of equation (15) on information known at time $t - 1$. Projecting equation (15) on information known at time t instead gives

$$E_tc_{t+1} = [1 + r(1 - B)]c_t - (1 + r)u_t + E_tu_{t+1}, \tag{23}$$

which is the time-series restriction tested in Hall's paper. Hall assumes that transitory consumption, u_t, is identically zero. Therefore, as formulated by Hall, the testable hypothesis implied by (23) is that other variables dated $t - 1$ or earlier have no additional explanatory value in predicting c_t if c_{t-1} is included in the regression. That is, equation (23) implies that

$$\begin{aligned}
&d_1 = [1 + r(1 - B)], \\
&d_i = 0, i = 2, \ldots, n \\
&c_i = 0, i = 1, \ldots, n
\end{aligned} \tag{24}$$

in the consumption autoregression (eq. [20]). The coefficient of lagged consumption, d_1, is estimated unrestricted. Presumably d_1 will be close to one.

The parameter values given in (24) satisfy the modified version of Sargent's cross-equation restrictions (eq. [22]): If the parameters of the consumption autoregression (c_i, d_i) are set equal to the values specified in (24), the equation restrictions stated in (22) all reduce to the form $0 \cdot a = 0$ or $0 \cdot b = 0$, so that the restrictions are satisfied for arbitrary values of the parameters of the income process (a_i, b_i). That is, imposing the restrictions $d_i = 0, i = 2, \ldots, n$; and $c_i = 0, i = 1, \ldots, n$, and estimating the income autoregression unconstrained would satisfy the restrictions in equation (22).

Note that Sargent's procedure yields restrictions in the form of a set of equations—one equation for each parameter of the consumption autoregression. In Hall's paper, however, the parameters of the consumption autoregression are constrained to particular values—lagged variables other than the first lagged value of consumption are constrained to enter with coefficients of zero. While correcting the definition of permanent income makes the two sets of restrictions consistent, in the sense that the set of parameter values in (24) satisfies the equation restrictions in (22), there remains the issue of why the two sets of restrictions are not identical in form.

The set of equation restrictions in (22) is derived by modeling $E_{t-1}c_t$ as a general (unrestricted) autoregressive process, then solving for the restrictions across the parameters of the consumption and income autoregressions which are implied by equation (16). However, since equation (23) is also an implication of the rational-expectations and permanent-income hypotheses, no additional assumptions are required to restrict the consumption autoregression to a first-order process. If the parameters of the consumption autoregression are assigned the values implied by equation (23), the equation restrictions in (22) all reduce to the form $0 = 0$. That is, once the restrictions are imposed within the consumption autoregression, no further restrictions are implied by the set of restrictions in (22). Thus the difference in form of the two sets of restrictions, (22) and (24), is attributable to the fact that, in deriving the restrictions in (22), $E_{t-1}c_t$ is modeled as an unnecessarily general autoregressive process.

At this point, the conflict between the Hall and Sargent papers concerning the restrictions on the systematic part of the consumption autoregression has been resolved in favor of Hall's formulation. The two authors also differ in their execution of the test, Sargent estimating the autoregressions with quasi-differenced data and Hall estimating the autoregression with the data in levels.

Referring to equation (15), note that the disturbance in the regression of c_t on c_{t-1} and other lagged variables will include a term which represents the revision in permanent income in period t as well as a moving average of transitory consumption. Using ξ_t to represent the revision in permanent income,

$$\xi_t = rB \sum_{s=0}^{\infty} \left(\frac{1}{1+r}\right)^{s+1} (E_t - E_{t-1}) x_{t+s}, \tag{25}$$

the consumption autoregression, under the null hypothesis, is:

$$c_t = [1 + r(1 - B)]c_{t-1} + \xi_t - (1 + r)u_{t-1} + u_t. \tag{26}$$

Hall assumes that transitory consumption, u_t, is identically zero. In his

analytical discussion, Sargent assumes u_t is an AR1 process, $u_t = \rho u_{t-1} + \omega_t$, and quasi-differences (26):

$$c_t - \rho c_{t-1} = [1 + r(1 - B)](c_{t-1} - \rho c_{t-2}) + \xi_t - \rho\xi_{t-1} + \omega_t - (1 + r)\omega_{t-1}. \tag{27}$$

In the econometric specification of the test, however, Sargent assumes that the disturbance in the consumption autoregression is uncorrelated with the regressors (p. 689). The disturbance will be uncorrelated with the regressors only if transitory consumption is identically zero, since ω_t is correlated with c_{t-1}. Quasi-differencing the data introduces a moving average of the revision in permanent income term, ξ_t. Since Sargent in effect assumes away transitory consumption by asserting that the error is uncorrelated with regressors, it seems more sensible to leave the regression in levels and thus preserve the white-noise property of the ξ_t error term.

The results of tests of the corrected version of the restrictions are presented in table 1. Sargent's Y_1 income variable was constructed as Y_1 = GNP − capital consumption − state, local, and federal tax receipts + transfers; his consumption variable C was total personal consumption expenditures. Both series are seasonally adjusted and measured in constant (1972) dollars. For comparison, table 1 also includes similar tests using Hall's income and consumption variables, which were per capita personal disposable income and per capita expenditures on services and nondurable goods. Hall's data were also seasonally adjusted and stated in 1972 dollars.[2] Following Sargent's specification, all regressions included a linear time trend, denoted t.

Sargent conducted his tests with the data in first differences, having assumed the autoregressive parameter in the transitory consumption process to be unity. Using differenced data, the test statistic for the corrected restrictions was 19.492, which indicates rejection of the hypothesis at the 99 percent confidence level but not at the 99.5 percent level. Testing the incorrect restrictions, Sargent reported (p. 696) a marginal confidence level of 1.000 when B, the MPC out of permanent income, was constrained to equal .90. The marginal confidence level of the test was only .697 when B was unrestricted; however, the resulting point estimate was "highly implausible" (6.06×10^7).

For the reasons stated above, I think that it is more appropriate to run the regressions in levels than in differences. With the data in

[2] While the data set used to rerun Hall's regressions was constructed according to his basic definitions of the variables, it is not an exact reproduction of his data set. The data set used in table 1 differs from Hall's because it incorporates subsequent NIPA revisions of the latest observations and because a different price deflator was used to deflate the income series

TABLE 1

Regression	Likelihood Ratio Statistic	Critical Value of $\chi^2(k)$	D-W	SER
		Sargent's Data		
Sargent's sample (1948:II–1972:IV)				
1. Differenced data				
Restricted ($\Delta c_t = \alpha + d_1 \Delta c_{t-1} + \gamma t$)			1.95	4.04
Unrestricted ($\Delta c_{t-i}, i = 2$–4 and $\Delta y_{t-i}, i = 1$–4 included)	19.492	$\chi^2_{.90}(7) = 18.48$	1.95	3.81
2. Data in levels				
Restricted ($c_t = \alpha + d_1 c_{t-1} + \gamma t$)			2.00	4.04
Unrestricted ($c_{t-i}, i = 2$–4 and $y_{t-i}, i = 1$–4 included)	17.054	$\chi^2_{.97.5}(7) = 16.01$	1.95	3.85
Hall's sample (1948:I–1977:I, data in levels)				
Restricted ($c_t = \alpha + d_1 c_{t-1} + \gamma t$)			1.63	4.82
Unrestricted ($c_{t-i}, i = 2$–4 and $y_{t-i}, i = 1$–4 included)	17.326	$\chi^2_{.97.5}(7) = 16.01$	1.95	4.62
Unrestricted ($c_{t-i}, i = 2$–4 included)	9.944	$\chi^2_{.97.5}(3) = 9.35$	1.97	4.68
Unrestricted ($y_{t-i}, i = 1$–4 included)	15.712	$\chi^2_{.90}(4) = 14.86$	2.01	4.59
		Hall's Data and Sample (1948:I–1977:I)		
Restricted ($c_t = \alpha + d_1 c_{t-1} + \gamma t$)			1.68	14.27
Unrestricted ($c_{t-i}, i = 2$–4 and $y_{t-i}, i = 1$–4 included)	12.04	$\chi^2_{.90}(7) = 12.02$	1.96	13.99
Unrestricted ($c_{t-i}, i = 2$–4 included)	5.004	$\chi^2_{.90}(3) = 6.25$	1.95	14.15
Unrestricted ($y_{t-i}, i = 1$–4 included)	10.044	$\chi^2_{.95}(4) = 9.49$	2.03	13.91

levels, the test statistic drops to 17.054, which is significant at the 97.5 percent level. The difference in sample periods does not affect the results much.

The results are sensitive to the definition of consumption, however. Using Hall's data, which omit expenditures on durable goods from the consumption variable, the test statistic drops to 12.04, which is barely significant at the 90 percent level. Thus the difference in the consumption variables used by the two authors accounts for a major part of the gap between Hall's results and those reported by Sargent. Correcting the restrictions and omitting the autoregressive transformation of the data does have a noticeable effect, though, decreasing the marginal confidence level from the 1.000 reported by Sargent to 97.5 percent.

III. The Role of Current Income in Signaling Changes in Permanent Income

The model in this section is based on the notion that, in forming their expectations of future income, rational economic agents will exploit the fact that income is a stochastic process which exhibits a high degree of serial correlation. Based on the observation that the current realization of income is greater than anticipated, agents will revise their expectations of future income, and therefore their permanent income, upward. Further, the magnitude of the revision in permanent income associated with the realization of a disturbance in current income will depend on the parameters of the time-series representation of current income. In this section, the quantitative relationship between an innovation in current income and the revision in permanent income is derived for a general ARMA representation of the income process.

This section uses the innovation in the income process to model the period-by-period revisions in permanent income and develop a simple structural econometric model of consumption. The implications of the permanent income–rational expectations hypothesis are tested in the context of the structural model. The Hall-Sargent test of the permanent income–rational expectations hypothesis is then shown to be a test based on the reduced form of the structural model developed here.

The Income Process

More formally, the time-series behavior of aggregate income is modeled as an ARMA process which is driven by a series of independent

shocks or "innovations" in the process.[3] Thus,

$$y_t = \rho_1 y_{t-1} + \rho_2 y_{t-2} + \ldots + \rho_p y_{t-p} + \epsilon_t + \phi_1 \epsilon_{t-1} + \phi_2 \epsilon_{t-2} + \ldots + \phi_q \epsilon_{t-q}, \quad (28)$$

or $y_t = [1 + M(L)]/[1 - A(L)]\epsilon_t$ where L is the lag, or backshift, operator; $A(L) = \rho_1 L + \rho_2 L^2 + \ldots + \rho_p L^p$; $M(L) = \phi_1 L + \phi_2 L^2 + \ldots + \phi_q L^q$; and ϵ is a white-noise disturbance. The income variable is expressed in deviations from trend.

An ARMA model of order (p, q) can be equivalently characterized as a pure moving average model of infinite order. That is, the time-series representation

$$y_t = \epsilon_t + \psi_1 \epsilon_{t-1} + \psi_2 \epsilon_{t-2} + \psi_3 \epsilon_{t-3} + \ldots \quad (29)$$

is numerically equivalent to the model

$$y_t = \rho_1 y_{t-1} + \rho_2 y_{t-2} + \ldots + \rho_p y_{t-p} + \epsilon_t + \phi_1 \epsilon_{t-1} + \phi_2 \epsilon_{t-2} + \ldots + \phi_q \epsilon_{t-q} \quad (30)$$

for $\psi_s = \phi_s + \Sigma_{j=1}^{s} \rho_j \psi_{s-j}$.

Expressed in the pure moving average form, the time-series model summarizes the relationship between the forecasted value of the variable and the realization of the disturbance term:

$$\frac{dy_{t+s}}{d\epsilon_t} = \frac{dy_t}{d\epsilon_{t-s}} = \psi_s. \quad (31)$$

Since the disturbance terms are white noise by construction, the expectation of all future values of ϵ is zero. When the current forecast error ϵ_t is realized, the predicted value of income at each date in the future is revised proportionally to ϵ_t:

$$E_t y_{t+s} - E_{t-1} y_{t+s} = \psi_s \epsilon_t. \quad (32)$$

The present value of the revision, at time t, in the anticipated path of future income is

$$\sum_{s=0}^{\infty} \left(\frac{1}{1+r}\right)^s (E_t y_{t+s} - E_{t-1} y_{t+s}) = \left[\sum_{s=0}^{\infty} \left(\frac{1}{1+r}\right)^s \psi_s\right] \epsilon_t. \quad (33)$$

[3] It should be emphasized that the paper uses the ARMA framework to quantify the extent to which income is serially correlated only because the framework is a simple way of describing the time-series behavior of income on the aggregate level. The treatment of aggregate income as a stable ARMA process is not meant to imply that individuals view income as pure time series in forming expectations about their future income. On the individual level all that is assumed is that consumers will exploit the serial correlation that characterizes their individual income streams. For instance, a student who is unable to find a summer job realizes that summer jobs will probably be scarce next year as well; the production worker who is currently laid off realizes that he may not be rehired for some weeks; the stockbroker whose commissions have declined in volume may expect his income to remain below normal for 4–6 quarters.

In the infinite horizon case, the value of $\Sigma_{s=0}^{\infty} [1/(1 + r)]^s \psi_s$ can be calculated directly from the parameters ϕ_s, ρ_j of the ARMA (p, q) representation of the process. It is shown in Appendix I that

$$\sum_{s=0}^{\infty} \left(\frac{1}{1 + r}\right)^s \psi_s = \frac{1 + \sum_{s=1}^{q} \left(\frac{1}{1 + r}\right)^s \phi_s}{1 - \sum_{j=1}^{p} \left(\frac{1}{1 + r}\right)^j \rho_j}. \tag{34}$$

Treatment of Unanticipated Capital Gains

In the simplest version of the model discussed in Section I, the return to real wealth, r, was assumed to be constant. This assumption implies that there are no unanticipated capital gains or losses on nonhuman wealth. In the absence of unanticipated capital gains, future labor income is the only component of permanent income which is not perfectly predictable. In this case the period-by-period revisions in permanent income would be modeled as proportional to the innovations in labor income alone.

As an empirical matter, however, unanticipated capital gains and losses on nonhuman wealth probably constitute a significant fraction of the revisions in permanent income which this model is trying to capture. The following discussion describes the construction of an approximate measure of unanticipated capital gains on nonhuman wealth.

Conceptually, the unanticipated capital gain on nonhuman wealth could be defined as the present value of the revision in the expected earnings associated with the current asset holdings:

$$\text{capital gain in period } t + 1 = \sum_{s=0}^{\infty} \left(\frac{1}{1 + r}\right)^{s+1} (E_{t+1} - E_t) r_{t+s} w_t, \tag{35}$$

where r_{t+s} is the rate of return to capital in period $t + s$. The actual path of nonlabor income, $r_{t+s} w_{t+s}$, will reflect both the "exogenous" movements in the return to capital, r_{t+s}, which we want to capture, and the endogenous changes in earnings flows which result from the individual's decisions to accumulate or decumulate nonhuman wealth, w_{t+s}. I assume that changes in the rate of return to capital, r_{t+s}, are quantitatively more important than the endogenous changes in w_{t+s} in determining the time-series properties of the observed path of nonlabor income. With this assumption, unanticipated capital gains are approximated as the present value of the revision in expected nonlabor income.

Repeating the derivation of the consumption equation as in Section I, the unanticipated capital gains in real wealth enter the equation symmetrically with the present value of the revision in expected labor income:

$$c_t = c_{t-1} + r\left[\sum_{s=0}^{\infty}\left(\frac{1}{1+r}\right)^{s+1}(E_t - E_{t-1})(x_{t+s} + y^k_{t+s})\right] - (1+r)u_{t-1} + u_t, \tag{36}$$

where x_{t+s} is future labor income and y^k_{t+s} is future nonlabor income.

Substituting into this equation the time-series characterization of the revision in the predicted path of future income yields:

$$c_t = c_{t-1} + r\left[\sum_{s=0}^{\infty}\left(\frac{1}{1+r}\right)^{s+1}\psi_s\right]\epsilon_t - (1+r)u_{t-1} + u_t, \tag{37}$$

where ψ_s are the parameters of the moving average representation, and ϵ_t the innovations, of the disposable income series, $y_t = x_t + y^k_t$. To simplify the notation, let $\Phi = \Sigma_{s=0}^{\infty}[1/(1+r)]^{s+1}\psi_s$. Using this notation, the present value of the revision, at time t, in the anticipated path of future income is denoted $\Phi\epsilon_t$; the revision in permanent income is denoted by $r\Phi\epsilon_t$. Equation (37) can be rewritten as:

$$c_t = c_{t-1} + r\Phi\epsilon_t - (1+r)u_{t-1} + u_t \tag{38}$$

Finally, it is necessary to consider the effect on aggregate consumption of a trend in per capita income. While an individual's permanent income can be defined so that its movement over time is trendless, the presence of a trend in per capita income will induce a trend in aggregate permanent income and therefore in aggregate consumption. If per capita income has a positive trend, due to increasing productivity, for example, later generations will have greater lifetime wealth than earlier ones. As the older generations die and are replaced by the younger generations, aggregate consumption will trend upward at the same rate of growth as per capita income. No change in the expectation of future income is involved in this trend movement of aggregate consumption, nor does any particular individual change his level of consumption. Since the model is intended to explain revisions in planned consumption which are caused by changes in the expectations about future income, equation (38) applies to the movement of consumption around a trend attributable to the trend in per capita income.

A basic empirical implication of the model developed in Section I is that, even if the behavioral marginal propensity to consume out of current income is zero, consumption should respond to innovations in current income because these innovations provide new information about future income and therefore induce revisions in permanent income. Expressing the income forecast error in terms of current and lagged income yields the following system of equations:

$$y_t = \mu_1 + \rho_1 y_{t-1} + \rho_2 y_{t-2} + \ldots + \rho_p y_{t-p} + \epsilon_{1t}, \tag{39}$$

$$\Delta c_t = \mu_2 + k\Phi(y_t - \mu_1 - \rho_1 y_{t-1} - \ldots - \rho_p y_{t-p})$$
$$+ \beta_0 \Delta y_t + \beta_1 \Delta y_{t-1} + \beta_2 \Delta y_{t-2} + \ldots + \beta_{p-1} \Delta y_{t-(p-1)} + \epsilon_{2t}, \tag{40}$$

where

$$\Phi = \left[\frac{1}{1 - \frac{\rho_1}{1+r} - \frac{\rho_2}{(1+r)^2} - \ldots - \frac{\rho_p}{(1+r)^p}} \right],$$

r is the quarterly interest rate, and k represents the annuity rate. (If the individual's time horizon is finite, the annuity rate will not be equal to the interest rate, as in the infinite horizon case.)

In the unrestricted version of the model, the first difference of consumption responds to current and lagged changes in income as well as the innovation in the income process. In equation (40) the β_i coefficients are measures of the "excess sensitivity" of consumption to current income, that is, sensitivity in excess of the response attributable to the new information contained in current income. The empirical work discussed in this section provides estimates of the β_i parameters and tests the implication of the permanent income–rational expectations hypothesis that consumption exhibits no excess sensitivity to current income, $\beta_i = 0$.

The permanent income–rational expectations hypothesis could be tested by estimating the system in its structural representation, equations (39) and (40), by full information maximum likelihood (FIML) both with and without the restriction $\beta_i = 0$ and forming a likelihood ratio statistic.

As discussed below, the unrestricted version of the system is just identified. Because the system is just identified, the maximum likelihood estimates of the structural parameters can be recovered from the estimated coefficients of the reduced form. Working with the reduced form has several advantages: First, estimation of the reduced form is computationally simpler than estimating the structural repre-

sentation by FIML; second, analysis of the reduced form clarifies the exact relationship between the likelihood ratio test of $\beta_i = 0$ proposed here and the F-test proposed by Hall. It will be shown that Hall's test is a test based on the reduced form of the consumption equation (40). Further, it will be shown that because the structural system (eqq. [39] and [40]) is just identified, the test statistic from Hall's test, based on the reduced form of the consumption equation alone, is numerically identical to the test statistic for the hypothesis $\beta_i = 0$ in the two-equation structural system. For the purpose of determining the statistical significance of a departure from the permanent income–rational expectations hypothesis, both Hall's test and the test proposed here yield identical answers. For the purposes of assessing whether or not an observed departure from the hypothesis is of quantitative significance, however, the test proposed here has the advantage of permitting the recovery of the point estimates of the structural parameters, the β_i, and their individual standard errors, whereas Hall's procedure estimates only reduced-form coefficients.

Identification of the Model

Before discussing the identification of the system, I will make explicit my assumptions concerning the properties of the income-forecast error, ϵ_1, and the disturbance in the consumption equation, ϵ_2.

Conceptually, one would like to estimate the 1-period-ahead forecast of income, *conditional on all available information*, $E(y_t|I_{t-1})$. Since the estimated prediction equation (39) includes only a very limited subset of the available information, the predictive value of all the lagged values of variables observed by the individual, but not explicitly incorporated in the regression, will be reflected in the error term, ϵ_1. Using the notation $R(L)x_{t-1}$ to represent the discrepancy between the 1-period-ahead forecast of income conditional on the complete information set I_{t-1} and the income forecast conditional on the limited information set actually included in the regression, $R(L)x_{t-1} = E(y_t|I_{t-1}) - E(y_t|y_{t-1}, \ldots, y_{t-p})$, the disturbance term ϵ_1 in equation (39) will be the sum of the "true" innovation $\tilde{\epsilon}_{1t}$ and the $R(L)x_{t-1}$ term,

$$\epsilon_{1t} = \tilde{\epsilon}_{1t} + R(L)x_{t-1}. \tag{41}$$

The term $R(L)x_{t-1}$ can be thought of as measurement error due to the omission of elements of the available information set from the forecasting equation. The $R(L)x_{t-1}$ term is assumed to be serially uncorrelated and uncorrelated with the explanatory variables included in the income-prediction equation. By definition, the "true" innovation $\tilde{\epsilon}_t$ is uncorrelated with the $R(L)x_{t-1}$ term as well as the included explanatory variables.

Since the "true" income innovation, $\tilde{\epsilon}_{1t} = y_t - \rho_1 y_{t-1} - \ldots - \rho_p y_{t-p} - R(L)x_{t-1}$, is the explanatory variable in the consumption equation, the consumption disturbance ϵ_{2t} will contain the term $-k\Phi R(L)x_{t-1}$, where $k\Phi$ is the coefficient of the true innovation.

The systematic part of the consumption equation models the response of consumption to the new information contained in the current observation on income. Obviously there are many sources of new information about future income in addition to the current observation on income itself. These omitted current variables which induce individuals to revise their expectations about future income will also contribute to the disturbance in the consumption equation. Let the variable θ_t represent the revision, in period t, in permanent income induced by other (nonincome) new information about future income. The "omitted news" disturbance θ_t is assumed to be serially uncorrelated and uncorrelated with lagged variables.

Because consumption is one component of current income, the omitted news term will have a nonzero correlation with the true innovation in income, $\tilde{\epsilon}_{1t}$. If, in period t, current variables other than income induce individuals to revise their forecasts of future income upward, θ_t will be positive. The positive realization of θ_t increases current consumption and therefore increases current income. This implies that the omitted news term θ_t is positively correlated with the true income innovation, $\tilde{\epsilon}_{1t}$.

In addition to the measurement error term, $-k\Phi R(L)x_{t-1}$, and the omitted news term, θ_t, the consumption equation will, in principle, also include a moving average of transitory consumption, $\eta_t - (1 + k)\eta_{t-1}$. For the purpose of estimating the model, however, the transitory consumption component of the error term will be ignored. Even though an individual's transitory consumption in a given quarter may be nonnegligible in relation to his permanent consumption, the per capita value of transitory consumption will be small relative to permanent per capita consumption if the individual realizations of transitory consumption are largely independent across the population. I think that it is reasonable to assume that the quantitative contribution of the transitory consumption term to the overall error ϵ_{2t} is small compared to the contribution of the measurement error term, $-k\Phi R(L)x_{t-1}$, and the omitted news error term, θ_t.

This discussion leads to the following characterization of the structural disturbances:

$$\epsilon_{1t} = R(L)x_{t-1} + \tilde{\epsilon}_{1t} \tag{49}$$

and

$$\epsilon_{2t} = -k\Phi R(L)x_{t-1} + \theta_t. \tag{43}$$

Based on these assumptions, ϵ_1 and ϵ_2 are contemporaneously correlated,

$$\text{cov}\,(\epsilon_{1t}, \epsilon_{2t}) = -k\Phi \text{ var } [R(L)x_{t-1}] + \text{cov}\,(\theta_t, \tilde{\epsilon}_{1t}), \tag{44}$$

but not serially correlated.

In order to simplify the discussion, identification of the system will be analyzed for a first-order AR income process. The generalization to a longer AR process is straightforward.

The unrestricted system is:

$$y_t = \mu_1 + \rho y_{t-1} + \epsilon_{1t} \tag{45}$$

and

$$\Delta c_t = \mu_2 + k\Phi\epsilon_{1t} + \beta\Delta y_t + \epsilon_{2t}, \tag{46}$$

where $\Phi = 1/\{1 - [\rho/(1 + r)]\}$. The value of the parameter k, which represents the marginal propensity to consume out of wealth, is imposed a priori. The parameter Φ, which represents the change in wealth associated with an innovation in income, is a function of the interest rate and the autoregressive parameters of the income process. The value of the interest rate is imposed a priori, so the identification of ρ identifies Φ as well.

Since $\epsilon_{1t} = y_t - \rho y_{t-1}$, the consumption equation can be rewritten as:

$$\Delta c_t = \mu_2 + (k\Phi + \beta)\epsilon_{1t} + \beta(\rho - 1)y_{t-1} + \epsilon_{2t}. \tag{47}$$

The covariances of the structural disturbances will be denoted as follows: var $(\epsilon_{1t}) = \sigma_1^2$; var $(\epsilon_{2t}) = \sigma_2^2$; and cov $(\epsilon_{1t}, \epsilon_{2t}) = \sigma_{12}$.

The parameters of the income process, μ_1, ρ, and σ_1^2, are identified by the income data alone:

$$\text{cov}\,(y_t, y_{t-1}) = \rho \text{ var}\,(y_t), \tag{48}$$

$$\text{var}\,(y_t) = \rho^2 \text{ var}\,(y_t) + \sigma_1^2, \tag{49}$$

and

$$\mu_1 = (1 - \rho)\mu_y, \tag{50}$$

where μ_y is the mean of y.

Using the fact that ϵ_{1t} is by construction orthogonal to y_t and ϵ_{2t} is assumed uncorrelated with y_{t-1},

$$\text{cov}\,(\Delta c_t, y_{t-1}) = \beta(\rho - 1) \text{ var } y_t, \tag{51}$$

$$\text{cov}\,(\Delta c_t, y_t) = \beta(\rho - 1) \text{ cov}\,(y_t, y_{t-1}) + (k\Phi + \beta)\sigma_1^2 + \sigma_{12}, \tag{52}$$

$$\text{var}\,(\Delta c_t) = \beta^2(\rho - 1)^2 \text{ var}\,(y_t) + (k\Phi + \beta)^2\sigma_1^2 + 2(k\Phi + \beta)\sigma_{12} + \sigma_2^2, \tag{53}$$

and

$$\mu_2 = \mu_{\Delta c_t} + \beta(1 - \rho)\mu_{y_t}, \tag{54}$$

where $\mu_{\Delta c_t}$ is the mean of Δc_t.

The model is just identified—equations (48) through (54) constitute a system of seven equations in seven unknown parameters. Because the model is just identified, indirect least squares—that is, estimating the model as a reduced form, then transforming the reduced-form coefficients to recover the structural parameters—is equivalent to FIML.

Consider the following reduced form:

$$y_t = \mu_1 + \rho y_{t-1} + \nu_{1t} \tag{55}$$

$$\Delta c_t = \mu_2 + \beta(\rho - 1)y_{t-1} + \nu_{2t}, \tag{56}$$

where $\nu_{1t} = \epsilon_{1t}$ and $\nu_{2t} = (k\Phi + \beta)\epsilon_{1t} + \epsilon_{2t}$. The covariance matrix of the reduced form, written in terms of the structural covariances, is:

$$\Omega = \begin{bmatrix} \sigma_1^2 & (k\Phi + \beta)\sigma_1^2 + \sigma_{12} \\ (k\Phi + \beta)\sigma_1^2 + \sigma_{12} & (k\Phi + \beta)^2\sigma_1^2 + 2(k\Phi + \beta)\sigma_{12} + \sigma_2^2 \end{bmatrix}. \tag{57}$$

Note that the maximum likelihood estimates of the structural coefficients—μ_1, μ_2, ρ, β—are independent of the value of $k\Phi$, the parameter which measures the predicted response of consumption to the innovation in current income. That is, the parameters μ_1, μ_2, ρ, and β are identified by equations (48), (49), (50), and (54), none of which contains $k\Phi$.

Examination of equations (53) and (54), or, equivalently, the reduced-form covariance matrix, equation (57), indicates that a priori restriction on the value of $k\Phi$ makes it possible to recover the structural covariances σ_{12} and σ_2^2 from the estimated covariances of the reduced form:

$$\text{cov}\,(\nu_{1t}, \nu_{2t}) = (k\Phi + \beta)\sigma_1^2 + \sigma_{12} \tag{58}$$

and

$$\text{var}\,(\nu_{2t}) = (k\Phi + \beta)^2\sigma_1^2 + 2(k\Phi + \beta)\sigma_{12} + \sigma_2^2. \tag{59}$$

Since indirect least squares and FIML are equivalent in the just-identified case, the likelihood ratio statistic for the test of the restriction $\beta = 0$ computed by estimating the structural form, equations (45) and (46), by FIML will be numerically identical to the likelihood ratio

test statistic for $\beta = 0$ computed by estimating the reduced form, equations (55) and (56), by multivariate least squares.[4]

Comparison of the Test of $\beta_1 = 0$ with Hall's Test

In the equation system actually estimated, income was modeled as an eighth-order pure autoregressive process:

$$y_t = \mu_1 + \rho_1 y_{t-1} + \rho_2 y_{t-2} + \ldots + \rho_8 y_{t-8} + \epsilon_{1t}, \tag{60}$$

$$\begin{aligned}\Delta c_t = \mu_2 + k\Phi(y_t - \mu_1 - \rho_1 y_{t-1} - \rho_2 y_{t-2} - \ldots - \rho_8 y_{t-8}) \\ + \beta_0 \Delta y_t + \beta_1 \Delta y_{t-1} + \ldots + \beta_7 \Delta y_{t-7} + \epsilon_{2t}.\end{aligned} \tag{61}$$

Like the simpler system discussed above, this system is just identified. If a less general alternative hypothesis were specified—for example, if only the contemporaneous and three lagged first differences in income were included in the unrestricted model—the system would be overidentified. The particular specification of the alternative hypothesis was deliberately chosen in order to achieve just identification. Because the model is just identified, the maximum-likelihood estimates of the structural parameters can be obtained by estimating the reduced form:

$$y_t = \mu_1 + \rho_1 y_{t-1} + \rho_2 y_{t-2} + \ldots + \rho_8 y_{t-8} + \nu_{1t} \tag{62}$$

and

$$\begin{aligned}\Delta c_t = (\mu_2 + \beta_0 \mu_1) + \beta_0[(\rho_1 - 1)y_{t-1} + \rho_2 y_{t-2} + \ldots + \rho_8 y_{t-8}] \\ + \beta_1 \Delta y_{t-1} + \beta_2 \Delta y_{t-2} + \ldots + \beta_7 \Delta y_{t-7} + \nu_{2t},\end{aligned} \tag{63}$$

where $\nu_{1t} = \epsilon_{1t}$ and $\nu_{2t} = (k\Phi + \beta_0)\epsilon_{1t} + \epsilon_{2t}$. Using the notation π_i to denote the reduced-form coefficient of y_{t-i},

$$\Delta c_t = \tilde{\mu}_2 + \pi_1 y_{t-1} + \pi_2 y_{t-2} + \ldots + \pi_8 y_{t-8} + \nu_{2t}. \tag{64}$$

Equation (63) makes very explicit the relationship between the reduced-form coefficients π_i and the underlying structural parameters:

[4] Eq. (51) suggests a simple instrumental variables estimator of the excess sensitivity parameter $\hat{\beta}$. Noting that $(\rho - 1)$ var y_t = cov $(\Delta y_t, y_{t-1})$, eq. (51) can be rewritten as

$$\text{cov}(\Delta c_t, y_{t-1}) = \beta \,\text{cov}(\Delta y_t, y_{t-1}) \tag{51'}$$

Thus $\hat{\beta}$ is identified as the ratio $[\text{cov}(\Delta c_t, y_{t-1})]/[\text{cov}(\Delta y_t, y_{t-1})]$. Writing the consumption equation with the income innovation in the error term, $\Delta c_t = \mu_2 + \beta \Delta y_t + (k\Phi\epsilon_{1t} + \epsilon_{2t})$, one could estimate β by using y_{t-1} as an instrument for Δy_t. Lagged income, y_{t-1}, is clearly a valid instrument for Δy_t, since it is correlated with Δy_t but uncorrelated with the composite error term, $(k\Phi\epsilon_{1t} + \epsilon_{2t})$. Using y_{t-1} as an instrument for Δy_t, the instrumental variables estimate of β is $[\text{cov}(\Delta c_t, y_{t-1})]/[\text{cov}(\Delta y_t, y_{t-1})]$, which is exactly the expression for β obtained from eq. (51).

$$
\begin{aligned}
\tilde{\mu}_2 &= \mu_2 + \beta_0\mu_1 \\
\pi_1 &= \beta_0(\rho_1 - 1) + \beta_1 \\
\pi_2 &= \beta_0\rho_2 - \beta_1 + \beta_2 \\
\pi_3 &= \beta_0\rho_3 - \beta_2 + \beta_3 \\
&\cdot \\
&\cdot \\
&\cdot \\
\pi_7 &= \beta_0\rho_7 - \beta_6 + \beta_7 \\
\pi_8 &= \beta_0\rho_8 - \beta_7.
\end{aligned}
\tag{65}
$$

Regardless of the time-series process describing income, the restriction that consumption exhibits no excess sensitivity to current income, $\beta_0 = \beta_1 = \ldots = \beta_7 = 0$, implies that the reduced-form coefficients π_i are all equal to zero.

The reduced form expresses the conditional expectation of the endogenous variable, Δc_t, given the predetermined variables, which in this case are the lagged values of income. Thus, the empirical content of the restriction that $\beta_0 = \beta_1 = \ldots = \beta_7 = 0$ is that the conditional expectation of Δc_t given lagged values of income is zero. This implication of the permanent income–rational expectations hypothesis is precisely the hypothesis tested by Hall. Quoting from Hall's paper (Hall 1978, p. 976), "The tests of the stochastic implications of the life cycle–permanent income hypothesis carried out in this paper all have the form of estimating a conditional expectation, $E(c_t | c_{t-1}, x_{t-1})$, where x_{t-1} is a vector of data known in period $t - 1$, and then testing the hypothesis that the conditional expectation is actually not a function of x_{t-1}."

The analytical approach taken by this paper was to formulate a general structural model in which consumption responded to both innovations in permanent income and changes in current income. The restrictions imposed by the permanent income–rational expectations hypothesis on the structural parameters, $\beta_i = 0$, were shown to imply the restriction that the reduced-form parameters π_i were equal to zero. Hall's approach to formulating a test of the hypothesis avoided the econometric specification of a structural consumption equation. Eschewing structural estimation altogether, Hall analyzed the restrictions imposed by the permanent income–rational expectations hypothesis on the conditional expectation of Δc_t given lagged variables. Thus Hall arrived at the restrictions imposed on the reduced-form consumption equation by an analytical shortcut, rather

than derive the reduced-form restrictions from an explicit structural econometric model, he exploited the interpretation of a reduced form as stating a conditional expectation and considered the implications of the permanent income–rational expectations hypothesis for the conditional expectation of consumption.

Although both Hall's analysis and the structural model developed in this paper lead to the same restrictions on the reduced form of the consumption equation, the test procedure used here differs from Hall's in that the restriction is tested by computing the likelihood ratio statistic for the two-equation reduced form (eqq. [62] and [63]), while Hall's test statistic is computed by estimating the consumption equation (eq. [64]) alone. As proved in Appendix II, however, the two tests yield numerically identical values of the likelihood ratio statistic for the hypothesis that the conditional expectation of consumption given lagged income is zero.[5]

The basic advantage of the structurally derived test over Hall's test is that it permits the recovery of the point estimates of the structural parameters measuring the excess sensitivity of consumption to current income, the β_i, and their individual standard errors. With Hall's test it is difficult to interpret the quantitative importance of a departure from the predicted behavior of consumption since only the reduced-form coefficients are estimated.

Data

The test of the restriction $\beta_i = 0$ is a test of the joint hypothesis that (*a*) consumers exploit the time-series properties of income in forming their estimates of permanent income and (*b*) the adjustment of consumption to a revision in permanent income is fully achieved within the quarter that the new information becomes available. In order to limit the consumption concept to a component which can be adjusted rapidly and smoothly to changes in permanent income, the consumption concept used as the dependent variable was expenditures on nondurable goods. Even if one went to the trouble of converting the National Income and Product Accounts (NIPA) data on expenditures on durable goods into a series of imputed service flows, the existence of substantial transaction costs of adjusting stocks of durable goods suggests that the consumption of services of durable goods would exhibit lagged adjustment to changes in permanent income. The

[5] I am indebted to Hall for suggesting that I estimate the model by indirect least squares instead of FIML. The observation that the test based on the two-equation reduced form and the test based on the reduced form of the consumption equation alone will yield numerically identical values of the likelihood ratio statistic is also due to Hall.

NIPA data on consumption of services are subject to a similar objection, since services are defined to include the imputed service flow from housing. Nondurable goods represent about 45 percent of total personal consumption expenditures.

Both the consumption series and the income series, which was disposable personal income, were seasonally adjusted and expressed in real per capita terms. After allowing for the construction of the lagged variables, the sample period used in the estimation was 1949:III–1979:I.

The income process was modeled as an eighth-order autoregressive process around an exponential trend. In order to simplify the estimation, the income data series was first detrended by its estimated exponential trend for the sample period 1947:I–1979:I of .00565259 per quarter, and the model was estimated with the detrended series. As mentioned above, the presence of a trend in per capita income will give rise to a trend in per capita consumption. Because the consumption variable is expenditures on nondurable goods rather than total consumption, the trend in nondurable goods consumption will not be equal to the trend in income unless the income elasticity of nondurables consumption is unity. For this reason, the consumption series was detrended by its own estimated trend, which was .0032616 per quarter. Both the income and consumption series were expressed in units of (constant) dollars per capita.

Empirical Results

The reduced form was written out in terms of the structural parameters and estimated by multivariate least squares. The unconstrained system estimated was:

$$y_t = \mu_1 + \rho_1 y_{t-1} + \rho_2 y_{t-2} + \ldots + \rho_8 y_{t-8} + \nu_{1t} \tag{66}$$

and

$$\begin{aligned}\Delta c_t = \tilde{\mu}_2 + \beta_0[(\rho_1 - 1)y_{t-1} + \rho_2 y_{t-2} + \ldots + \rho_8 y_{t-8}] + \beta_1 \Delta y_{t-1} \\ + \beta_2 \Delta y_{t-2} + \ldots + \beta_7 \Delta y_{t-7} + \nu_{2t}.\end{aligned} \tag{67}$$

The point estimates and their standard errors are reported in table 2A.

Imposing the restriction $\beta_0 = \beta_1 = \ldots = \beta_7 = 0$, the constrained system consisting of the income equation (66) and the consumption equation

$$\Delta c_t = \tilde{\mu}_2 + \nu_{2t} \tag{68}$$

was also estimated by multivariate least squares. The point estimates

TABLE 2

A. POINT ESTIMATES AND STANDARD ERRORS FOR EQUATIONS (66) AND (67)

μ_1	ρ_1	ρ_2	ρ_3	ρ_4	ρ_5	ρ_6	ρ_7	ρ_8
92.54	.964	.069	.074	−.360	.112	.157	−.008	−.053
(64.26)	(.092)	(.126)	(.125)	(.122)	(.123)	(.124)	(.122)	(.085)
$\bar{\mu}_2$	β_0	β_1	β_2	β_3	β_4	β_5	β_6	β_7
32.71	.355	.071	.049	−.116	.114	−.073	.050	−.012
(24.95)	(.275)	(.036)	(.038)	(.050)	(.071)	(.046)	(.036)	(.035)

B. POINT ESTIMATES AND STANDARD ERRORS FOR EQUATION (68)

μ_1	ρ_1	ρ_2	ρ_3	ρ_4	ρ_5	ρ_6	ρ_7	ρ_8	$\bar{\mu}_2$
55.62	.898	.067	.231	−.476	.280	−.045	.065	−.045	.068
(57.72)	(.082)	(.113)	(.112)	(.110)	(.111)	(.111)	(.110)	(.076)	(.862)

C. REDUCED-FORM COEFFICIENTS AND STANDARD ERRORS FOR EQUATION (69)

$\bar{\mu}_2$	π_1	π_2	π_3	π_4	π_5	π_6	π_7	π_8
32.71	.058	.002	−.138	.103	−.148	.179	−.065	.007
(25.95)	(.037)	(.051)	(.050)	(.049)	(.050)	(.050)	(.049)	(.034)

and standard errors, based on the constrained estimation, are reported in table 2B.

Using the notation L_c and L_u to denote the log likelihood of the constrained and unconstrained systems, respectively, the likelihood ratio statistic $\lambda = -2(L_c - L_u)$ for the hypothesis $\beta_0 = \beta_1 = \ldots = \beta_7 = 0$ is distributed $\chi^2(8)$. The value of the likelihood ratio test statistic, λ, was 27.024. Since the critical value of the $\chi^2(8)$ distribution is 21.96 for the 0.5 percent significance level, a test statistic of 27.0 represents a very decisive rejection of the hypothesis.

For comparison, Hall's version of the test using the reduced form of the consumption equation alone,

$$\Delta c_t = \bar{\mu}_2 + \pi_1 y_{t-1} + \pi_2 y_{t-2} + \ldots + \pi_8 y_{t-8} + \nu_{1t}, \tag{69}$$

was also estimated. The reduced-form coefficients and their standard errors are reported in table 2C.[6] The likelihood ratio statistic for the hypothesis that $\beta_0 = \beta_1 = \ldots = \beta_7 = 0$, it is necessary to compute the computer work confirms the assertion that the test of $\beta_0 = \beta_1 = \ldots =$

[6] The estimated constant term in the constrained version of eq. (69) was .068, with a standard error of .866.

TABLE 3

STANDARD ERRORS OF REGRESSION AND DURBIN-WATSON STATISTICS

	System Estimated					
	Two-Equation Reduced Form				Single-Equation Reduced Form	
	Unrestricted		Restricted		Unrestricted	Restricted
	y	Δc	y	Δc	Δc	Δc
SER ...	$22.50	$8.73	$23.01	$9.45	$8.73	$9.45
D-W ..	2.00	2.15	1.90	2.13	2.15	2.13

NOTE: y and Δc = dependent variables

$\beta_7 = 0$ in the context of the two-equation reduced form (eqq. [66] and [67]) and the test of $\pi_1 = \pi_2 = \ldots = \pi_8 = 0$ in the context of the reduced-form consumption equation alone, equation (69), yield numerically identical values of the likelihood ratio statistic.

The standard errors of regression (SER) and Durbin-Watson statistics are reported in table 3. The SER of each equation was computed by dividing the sum of squared residuals by the degrees of freedom rather than the number of observations. The standard errors have the interpretation of the dollars per capita prediction error of each equation. For comparison, the standard deviations of y and Δc were $71.57 and $9.45, respectively.

The discussion of the identification of the system showed that the maximum-likelihood estimates of the β_i parameters were independent of the value of $k\Phi$, the parameter which measures the predicted response of consumption to an innovation in current income. While the identification of the parameter $k\Phi$ is not required to test the hypothesis that $\beta_0 = \beta_1 = \ldots = \beta_7 = 0$, it is necessary to compute the value of $k\Phi$ in order to recover the parameters of the covariance matrix of structural disturbances.

By substituting the estimated parameters of the income process into equation (34), one can compute the value of Φ, which gives the capitalized value of the revision in expected future income or the change in wealth associated with an innovation in current income. For the estimated parameters of the income process, and an annual interest rate of 5 percent, an innovation in current income implies a change in wealth of 17.80 times the income innovation.[7] Assuming an

[7] In order to simplify the estimation, the income process was modeled as an 8-quarter pure AR process. Formal tests of more general mixed ARMA models of the income process identified the (3, 0, 4) specification (three AR parameters and four MA parameters) and the (1, 0, 4) specification as the most parsimonious mixed ARMA

average horizon of 20 years, the annuity rate is 0.0198 per quarter.[8] The change in permanent income associated with an innovation in current income is (0.0198)(17.80), or 0.35 times the income innovation. Since the consumption variable used to estimate the system was nondurables consumption rather than total consumption, the parameter $k\Phi$ will be the product of the marginal propensity to consume nondurables out of permanent income and the change in permanent income induced by an innovation in current income. With nondurable consumption accounting for 47.5 percent of total consumption, $k\Phi = (0.475)(0.0198)(17.80) = 0.1675$.

The estimated covariance matrix of the disturbances of the unrestricted reduced form is:

$$\hat{\Omega} = \begin{bmatrix} 467.808 & 79.8921 \\ 79.8921 & 70.5175 \end{bmatrix}.$$

Having identified the value of $k\Phi$ as 0.1675 and the value of β_0 as .355, the covariance matrix of the structural disturbances can be recovered using equation (57). The estimated covariance matrix of the structural disturbances is:

$$\hat{\Sigma} = \begin{bmatrix} 467.808 & -164.538 \\ -164.538 & 114.745 \end{bmatrix}.$$

Note that the covariance between the structural disturbances of the two equations, cov (ϵ_1, ϵ_2), is negative and very large relative to the variance of the disturbance in the consumption equation, ϵ_2. In the discussion of the properties of the structural disturbances, it was pointed out that the omission of elements of the available information set from the income-forecasting equation implies that the income residual is subject to measurement error and that the presence of measurement error gives rise to a negative covariance between the structural disturbances. The large negative estimate of cov (ϵ_1, ϵ_2) suggests that the measurement error introduced by using the income-forecast error as a measure of the "true" income innovation cannot be dismissed as negligible. The finding that the measurement error is quantitatively important indicates that attempts to estimate $k\Phi$

representations of the income process. The value of Φ implied by the (3, 0, 4) model was 16.41, by the (1, 0, 4) model, 15.45. Since the ARMA models identified through formal time-series analysis of the income series yield values of Φ very similar to the value of Φ computed from the 8-quarter AR process, there is no indication that the assumed pure AR process is inadequate in capturing the time-series properties of income.

[8] The quarterly annuity rate was calculated with the following formula: annuity rate $= r/(1 - e^{-rT})$. The quarterly interest rate, r, was assumed to be 0.0125, the horizon of the representative individual, T, was assumed to be 80 quarters.

directly by doing regressions of the change in consumption on the income residuals will be subject to severe downward bias.

The discrepancy between the predicted and observed behavior of consumption is quantitatively large as well as statistically significant. Since the dependent variable is consumption of nondurable goods rather than total consumption, the β_i coefficients represent the excess sensitivity of nondurables consumption alone. Considering that nondurable goods consumption represents 45 percent of total consumption expenditures, the point estimates of β_0, β_1, and β_2 of .355, .071, and .049, respectively, indicate a strong "excess" response of consumption to current income. Unfortunately, the estimate of the parameter β_0 is imprecise, as indicated by its fairly large standard error of .275.

Because of the large sampling error of the estimate of the crucial parameter β_0, these tests stop short of providing conclusive evidence that the permanent income–rational expectations hypothesis fails in a quantitatively significant way. Still, the large point estimate β_0, together with the statistical rejection at the 0.5 percent level of the zero restrictions on the β_i coefficients, constitute significant, if not conclusive, evidence against the hypothesis.

In Hall's paper, tests of the predictive value of lagged income did not reveal strong evidence against the permanent-income hypothesis.[9] Hall concluded, "The results of this paper have the strong implication that beyond the next few quarters consumption should be treated as an exogenous variable. . . . With respect to the analysis of stabilization policy, the findings of this paper [support] the view that policy affects consumption only as much as it affects permanent income" (p. 986).

Why is it that the tests reported here reveal substantial evidence against the permanent-income hypothesis, whereas Hall's did not? As shown above, the test based on the two-equation reduced form is not intrinsically more powerful than Hall's test. The structurally derived reduced-form test provides stronger results for two reasons, one obvious and the other more subtle. First, by permitting the recovery of the structural parameter estimates measuring the excess sensitivity of consumption to current income, the β_i, the structurally derived reduced-form test makes it possible to assess the quantitative significance of a departure from the null hypothesis.

Of the regressions reported by Hall, the regression of current consumption (of nondurable goods and services) on lagged consumption and four lagged values of income is closest to the reduced

[9] Hall was able to statistically reject the permanent-income hypothesis based on tests of the predictive value of a stock price index.

form estimated in this paper. On page 983 Hall reports the following coefficient estimates and standard errors.

$$\underset{(11)}{c_t = -23} + \underset{(.047)}{1.076c_{t-1}} + \underset{(.043)}{.049y_{t-1}} - \underset{(.052)}{.051y_{t-2}} - \underset{(.051)}{.023y_{t-3}} - \underset{(.037)}{.024y_{t-4}}. \tag{70}$$

Testing the hypothesis that the coefficients of all four lagged values of income were zero, Hall got an F-statistic of 2.0, which was high enough to reject the hypothesis at the 10 percent level, although not the 5 percent level. Noting that the first lagged value of income had a small positive coefficient and that the sum of the coefficients was small and negative, Hall concluded, "There is a statistically marginal and numerically small relation between consumption and very recent levels of disposable income" (p. 984).

The statement that the relationship between consumption and disposable income is "numerically small" refers to the magnitude of the reduced-form coefficients, the π_i, not the structural parameters β_i. Hall is explicit on the point that his regressions did not provide estimates of structural parameters: "No claim is made that the true structural relation between consumption and its determinants is revealed by this approach" (p. 977).

The obvious advantage of the structurally derived reduced-form test is that it permits the estimation of the structural parameters. In equation system (54), the reduced-form coefficients are written out in terms of the underlying structural parameters. Equation system (68) indicates that even if consumption responds strongly to current and lagged changes in income, that is, the β_i are large and positive, the reduced-form parameters will be small. Consider in particular the reduced-form coefficient of y_{t-1}:

$$\pi_1 = \beta_0(\rho_1 - 1) + \beta_1. \tag{71}$$

When nondurables consumption alone is used to form the dependent variable, the estimate of β_0 is .355 and the estimate of β_1 is .071. However, since the first autoregressive parameter in the income process, ρ_1, is estimated to be .964, the value of π_1 is only .058. When the consumption concept is defined as nondurable goods and services, β_0 and β_1 are estimated to be .461 and .128, respectively. The resulting reduced-form coefficient π_1 in this case is .112. In addition to the 4-quarter distributed lag, Hall reported the results for a test using a 12-quarter Almon lag. In both cases the sum of the reduced-form parameters was small and negative. Using equation system (65), note that

$$\sum_{i=1}^{8} \pi_i = \beta_0 \left(\sum_{i=1}^{8} \rho_i - 1\right). \tag{72}$$

The sum of the autoregressive parameters of the income process is about .95 for quarterly data. Therefore, even if β_0 has a healthy positive value, the sum of the reduced-form parameters will be small and negative.

While the test based on the reduced form of the consumption equation alone gives the same test statistic as the test based on the two-equation reduced form, it does not by itself provide any information about the quantitative importance of a departure from the null hypothesis. Observing that the reduced-form coefficients are small does not imply that the structural coefficients are small. To the contrary, large positive values of the β_i coefficients are associated with reduced-form parameters which are small and which sum to a negative value.

The second, more subtle advantage of the structural approach is as follows: Because it requires the explicit specification of a structural model, the structural approach provides some guidance concerning the specification of the reduced form.

While Hall's approach of analyzing the restrictions on the conditional expectation of consumption led to the correct conclusion that the reduced-form coefficients of lagged variables should be zero, it did not suggest any criteria for selecting a particular collection of variables to be included in the reduced form. By studying the structural specification, one learns that, under the alternative hypothesis ($\beta_0 > 0$), any lagged value of income which has a nonzero coefficient in the income autoregression will in general have a nonzero coefficient in the consumption reduced form.

Further, if one believes that the income process is best modeled as an AR process around an exponential trend, then the consumption reduced form should be fit to the detrended income series rather than the raw income series. Alternatively, one could leave the income series in its original (not detrended) form and include an exponential trend term, $e^{\gamma t}$, in the consumption reduced form. In Hall's original tests, the income series was not detrended and the reduced form did not include a trend term. If the income process does include a trend, Hall's tests are misspecified under the alternative hypothesis. One would expect that the failure to allow for the trend in income, which causes the reduced form to be misspecified under the alternative hypothesis, would reduce the power of the test. As an empirical matter, the treatment of the income trend does have a noticeable effect on the value of the likelihood-ratio statistic.[10] In all of the tests reported above, the income series have been detrended.

[10] In a comment on Hall's paper, Startz (1979) pointed out that the inclusion of an exponential trend in the regression tended to increase the value of the test statistic

Having shown that the test based on the bivariate reduced form and Hall's test based on the single-equation reduced form yield numerically identical values of the likelihood ratio statistic for comparable specifications, it is necessary to determine why the rational expectations–permanent income hypothesis is rejected at the 0.5 percent level in the tests reported here, while the hypothesis could not be rejected at the 5 percent level in Hall's original tests. In particular, one would like to know whether the difference in the test statistic is attributable to differences in the data sets, or to differences in the specification of the alternative hypothesis, in terms of the number of lagged values of income included and the treatment of the trend in income.

Hall used a broader concept of consumption—expenditures on nondurable goods and services—than the consumption concept, nondurable goods, used in this paper. Data on real per capita consumption of nondurables and services were constructed to provide a direct comparison to Hall's original results. While the data set used to rerun Hall's regressions was constructed according to his basic definitions of the variables, it is not an exact reproduction of his data set. The data set used here differs from Hall's because it incorporates subsequent NIPA revisions of the most recent observations and because a different price deflator was used to deflate the income series. In the regression of current consumption on lagged consumption and four lagged values of income, Hall obtained an F-statistic for the exclusion of the lagged values of income of 2.0, which is somewhat below the critical value of 2.45 for the 5 percent level. For the same sample period and specification, the test statistic was 2.75 using the reconstructed data set. Thus, minor differences in the construction of the data raise the significance level of the test statistic from somewhat below to somewhat above the 5 percent level. Consideration of the structural model indicates that the reduced form should include as many lagged values of income as the income autoregression and that some allowance for a trend in income should be made, either by detrending the income series or by including a trend term in the consumption reduced form. Using the original (not detrended) income series but increasing the number of lagged values of income included in the consumption reduced form from four to eight increased the value of the likelihood-ratio statistic from 10.95 to 22.616. This raises the significance level of the test from the 5 percent level to the 0.5 percent level. Detrending the income series also has a substantial effect on the test statistic. For the specification with four lagged values of income, the test statistic was 16.192 for the detrended income series (which is significant at the 0.5 percent level) as opposed to 10.95 for the raw income series. For the specification with eight

lagged values of income, detrending the income series raises the test statistic from 22.616 to 27.108. Thus, the insight concerning the appropriate specification of the reduced form which is obtained by explicitly formulating the underlying structural model was responsible for raising the significance level of the test from a marginal rejection to a decisive rejection.

IV. Conclusions

This paper develops a simple structural econometric model of consumption. In the general specification of the model, consumption responds to the changes in permanent income signaled by innovations in the current income process and to changes in current income itself. The response of consumption to current income beyond that attributable to the role of current income in signaling changes in permanent income is termed the "excess sensitivity" of consumption to current income. The econometric model is then used to estimate the excess sensitivity of consumption to current income and to test the implication of the permanent-income hypothesis that the excess sensitivity of consumption is zero.

The tests reveal substantial evidence against the permanent-income hypothesis. Using either nondurables consumption or consumption of nondurables and services as the dependent variable, the hypothesis that consumption exhibits no excess sensitivity to current income can be rejected at the 0.5 percent level. Using expenditures on nondurable goods as the consumption variable, the point estimate of the excess sensitivity of consumption to the contemporaneous change in income is .355. Keeping in mind that nondurable goods represent only a fraction of total personal consumption expenditures—about 45 percent—a point estimate of .355 represents a large departure from the permanent-income hypothesis.

The paper also shows that the test of the permanent income–rational expectations hypothesis proposed by Hall (1978) can be thought of as a test based on the reduced form of the structural model developed here. Rather than develop a structural econometric model and derive the associated restrictions on the reduced-form parameters, Hall's approach was to analyze the implications of the permanent income–rational expectations hypothesis for the conditional expectation of consumption given lagged variables. Since a reduced form has the interpretation of expressing a conditional expectation, Hall's analysis led directly to testable restrictions on the parameters of a reduced-form consumption equation. When the structural model is

just identified, both tests yield numerically identical values of the test statistic. While the structural approach does not provide a statistically more powerful test, in general, than the reduced-form approach, it does have the advantage of providing estimates of the structural parameters of the model, including estimates of the "excess sensitivity" of consumption to current income.

Appendix I

For the general model

$$y_t = \rho_1 y_{t-1} + \rho_2 y_{t-2} + \ldots + \rho_p y_{t-p} + \epsilon_t + \phi_1 \epsilon_{t-1} + \phi_2 \epsilon_{t-2} + \ldots + \phi_q \epsilon_{t-q} \quad \text{(A1)}$$

$$\sum_{s=0}^{\infty} \frac{dy_{t+s}}{d\epsilon_t} = \sum_{i=0}^{\infty} \psi_i, \quad \text{(A2)}$$

where the ψ_i are the infinite series of moving average parameters in the pure moving average representation of the model: $y_t = \epsilon_t + \psi_1 \epsilon_{t-1} + \psi_2 \epsilon_{t-2} + \ldots$. Solving for ψ_i in terms of the estimated ρ_i and ϕ_i,

$$\psi_0 = \phi_0 = 1 \quad \text{(A3)}$$

$$\frac{dy_t}{d\epsilon_{t-i}} = \frac{\partial y_t}{\partial \epsilon_{t-i}} + \sum_{j=1}^{i} \frac{\partial y_t}{\partial y_{t-j}} \frac{dy_{t-j}}{d\epsilon_{t-i}} \quad \text{for } i = 1, \infty \quad \text{(A4)}$$

$$\psi_i = \phi_i + \sum_{j=1}^{i} \rho_j \psi_{i-j} \quad \text{(A5)}$$

where $\rho_j = 0$ for $j > p$, $\phi_i = 0$ for $i > q$.

$$\sum_{i=1}^{\infty} \psi_i = \sum_{i=1}^{\infty} \phi_i + \sum_{i=1}^{\infty} \sum_{j=1}^{i} \rho_j \psi_{i-j} \quad \text{(A6)}$$

$$\text{adding } \psi_0 = \phi_0. \quad \text{(A7)}$$

$$\sum_{i=0}^{\infty} \psi_i = \sum_{i=0}^{\infty} \phi_i + \sum_{i=1}^{\infty} \sum_{j=1}^{i} \rho_j \psi_{i-j} \quad \text{(A8)}$$

$$\sum_{i=0}^{\infty} \psi_i = \sum_{i=0}^{\infty} \phi_i + \sum_{j=1}^{\infty} \rho_j \sum_{i=j}^{\infty} \psi_{i-j} \quad \text{(A9)}$$

$$\sum_{i=0}^{\infty} \psi_i = \sum_{i=0}^{\infty} \phi_i + \sum_{j=1}^{\infty} \rho_j \sum_{i=0}^{\infty} \psi_i \quad \text{(A10)}$$

$$\left(1 - \sum_{j=1}^{\infty} \rho_j\right) \sum_{i=0}^{\infty} \psi_i = \sum_{i=0}^{\infty} \phi_i \quad \text{(A11)}$$

$$\sum_{i=0}^{\infty} \psi_i = \frac{\sum_{i=0}^{\infty} \phi_i}{1 - \sum_{j=1}^{\infty} \rho_j} = \frac{1 + \sum_{i=1}^{q} \phi_i}{1 - \sum_{j=1}^{p} \rho_j}. \quad \text{(A12)}$$

Appendix II

Proof that the likelihood ratio statistic for the test of $\gamma = 0$ in the equation

$$\Delta c_t = \mu + \gamma y_{t-1} + \nu_t \tag{A13}$$

is numerically identical to the likelihood ratio statistic for the test of $\beta = 0$ in the bivariate reduced form:

$$\underset{(T \times 2)}{[y_t \quad \Delta c_t]} = \underset{(T \times 2)}{[1 \quad y_{t-1}]} \underset{(2 \times 2)}{\begin{bmatrix} \mu_1 & \mu_2 \\ \rho & \beta(\rho - 1) \end{bmatrix}} + \underset{(T \times 2)}{[\nu_{1t} \quad \nu_{2t}]}. \tag{A14}$$

The GLS transformation of equation (A14) is obtained by postmultiplying by the lower triangular matrix P such that $PP' = \Omega^{-1}$, where Ω is the covariance matrix of the reduced form:

$$P = \begin{bmatrix} P_{11} & 0 \\ P_{12} & P_{22} \end{bmatrix}.$$

It will be convenient to normalize P such that the lower-right-hand element is equal to one. Define

$$\tilde{P} = \begin{bmatrix} \tilde{P}_{11} & 0 \\ \tilde{P}_{12} & 1 \end{bmatrix}$$

where $\tilde{P}_{11} = (P_{11}/P_{22})$ and $\tilde{P}_{12} = (P_{12}/P_{22})$. Transforming (A14) by $\tilde{P}$ yields:

$$\tilde{P}_{11} y_t + \tilde{P}_{12} \Delta c_t = (\tilde{P}_{11}\mu_1 + \tilde{P}_{12}\mu_2) + y_{t-1}[\tilde{P}_{11}\rho + \tilde{P}_{12}\beta(\rho - 1)] + \tilde{\nu}_{1t} \tag{A15}$$

and

$$\Delta c_t = \mu_2 + y_{t-1}\beta(\rho - 1) + \nu_{2t} \tag{A16}$$

where $\tilde{\nu}_{1t} = \tilde{P}_{11}\nu_{1t} + \tilde{P}_{12}\nu_{2t}$.

Let L^1 denote the log likelihood of the transformed income equation (A15), L^2 denote the log likelihood of the consumption equation (A16), and $L^{1,2}$ denote the log likelihood of the bivariate system. Since cov $(\tilde{\nu}_{1t}, \nu_{2t}) = 0$, the value of the log likelihood function of the bivariate system is the sum of the values of the log likelihood of the two individual equations. $L^{1,2} = L^1 + L^2$. Imposing the constraint that $\beta = 0$ reduces the value of L^2, $L^2_c \leq L^2_u$. Imposing the restriction that $\beta = 0$ does not restrict the log likelihood function of the transformed income equation, since μ_1 and ρ are unrestricted parameters, $L^1_c = L^1_u$. Therefore, all of the decrease in the likelihood function of the bivariate system is attributable to the decrease in the likelihood function of the consumption equation alone: $L^{1,2}_u - L^{1,2}_c = L^2_u - L^2_c$.

References

Abel, Andrew, and Mishkin, Frederic S. "On the Econometric Testing of Rationality and Market Efficiency." Report 7933, Univ. Chicago, Center Math. Studies Bus. and Econ., August 1979.

Ando, Albert, and Modigliani, Franco. "The 'Life Cycle' Hypothesis of Saving: Aggregate Implications and Tests." *A.E.R.* 53 (March 1963): 55–84.

Barro, Robert J. "Permanent Income and Expected Future Income in a Time Series Study of Consumer Expenditure." Manuscript. Rochester, N.Y. Univ. Rochester, April 1978.

Bilson, John F. O. "The Rational Expectations Approach to the Consumption Function: A Multi-Country Study." Paper presented at the International Seminar on Macroeconomics, Paris, September 1979.

Bilson, John F. O., and Glassman, James E. "A Consumption Function with Rational Forecasts of Permanent Income." Manuscript. Chicago. Univ. Chicago, February 1979.

Blinder, Alan S. "Temporary Income Taxes and Consumer Spending." *J.P.E.* 89, no. 1 (February 1981): 26–53.

Friedman, Milton. *A Theory of the Consumption Function.* Princeton, N.J.: Princeton Univ. Press (for Nat. Bur. Econ. Res.), 1957.

Hall, Robert E. "Stochastic Implications of the Life Cycle–Permanent Income Hypothesis: Theory and Evidence." *J.P.E.* 86, no. 6 (December 1978): 971–87.

Hayashi, Fumio. "Estimation of Permanent Income Consumption Functions under Rational Expectations." Manuscript. Evanston, Ill., Northwestern Univ., April 1979.

Muth, John. "Optimal Properties of Exponentially Weighted Forecasts." *J. American Stats. Assoc.* 55 (June 1960): 299–306.

Sargent, Thomas J. "Rational Expectations, Econometric Exogeneity, and Consumption." *J.P.E.* 86, no. 4 (August 1978): 673–700.

Startz, Richard. "Comment on 'Stochastic Implications of the Life Cycle–Permanent Income Hypothesis'." Unpublished paper. Philadelphia. Univ. Pennsylvania, Wharton School, February 1979.

Measurement of Monopoly Behavior: An Application to the Cigarette Industry

Daniel A. Sumner

North Carolina State University

A simple scheme is proposed to measure monopoly pricing behavior. The coefficient of the tax rate term in a price equation identifies the ratio of price to marginal cost. No direct measurement of costs is required, so a major problem for other empirical studies of monopoly is avoided. Empirical results for cross-section time-series data support rejection of atomistic competition but also provide evidence against the operation of an effective cartel in the cigarette industry. The model represents an alternative interpretation of related results in a recent paper by Barzel. Application of the methodology to other markets is feasible.

I. Introduction

The cigarette industry is often used as an example of a less than perfectly competitive industry. For example, in the famous Tobacco Case of 1946, the major cigarette manufacturers were accused of operating an illegal cartel; though they were convicted, the consensus was that company behavior was not altered by the verdict.[1] This paper

This is a revised version of "Interstate Cigarette Price Differentials and the Measurement of Firm Level Demand Elasticities" (October 1979). I wish to thank Yoram Barzel, Robert Clark, James Easley, David Flath, Charles Knoeber, Michael Kusnic, Thomas MaCurdy, Michael McElroy, Ronald Schrimper, Michael Wohlgenant, and an unknown referee for their comments. This paper is no. 6626 of the Journal Series of the North Carolina Agricultural Research Service, Raleigh.

[1] See Nicholls (1949), Bain (1968), or Tennant (1971) for discussion of the 1946 case and of the "structure, conduct and performance" of the cigarette industry. Stigler (1966, p. 244) presents a table of industries with high concentration; cigarettes rank fourth in the share of assets held by the "giant firms." Friedman (1953, p. 37) also

[*Journal of Political Economy*, 1981, vol. 89, no. 5]
© 1981 by The University of Chicago. 0022-3808/81/8905-0009$01.50

attempts to investigate the alleged market power of cigarette manufacturers with a simple scheme for measuring the weighted-average price elasticity of demand facing firms in this industry.

The empirical examination of the pricing behavior of the cigarette industry finds an absence of evidence of an effective cartel, but the hypothesis of atomistic competition is also rejected. The approach uses data on excise taxes, but it does not depend on measuring the cost of production. Since reliable cost data are not generally available, this paper avoids an important weakness of directly measuring monopoly power by the ratio of product price to marginal cost. The model suggested here is also applicable to other industries with variation in excise taxes or some other directly measurable components of marginal cost.[2]

II. A Simple Model

The equilibrium output condition, that marginal revenue equals marginal cost, implies that

$$P_j = [\eta_j/(\eta_j + 1)]M_j, \tag{1}$$

where P_j is the product price, M_j is the marginal cost of production, and η_j is the price elasticity of the demand curve facing firm j. If the industry is purely monopolistic, η_j is equal to the market elasticity η^m; while if it is perfectly competitive, $\eta_j = -\infty$ and is independent of η^m. Thus η_j is a measure of the degree to which the firm is not strictly a price taker.[3] However, since good measurements of M_j are difficult to find, equation (1) does not offer a practical solution to the problem of determining potential monopoly power.

For a product like cigarettes, which has a per unit excise tax, the marginal cost term may be expanded as

$$M_j = X_j\beta + \epsilon_j + R, \tag{2}$$

discusses potential monopoly power of cigarette firms, predicting that in an analysis of response to tax rate changes, "broadly correct results will be obtained by treating cigarette firms as if they were producing an identical product and were in perfect competition."

[2] The empirical model derived in this paper is similar to that used for a different purpose by Barzel (1976) (and also in a comment by Johnson [1978]). The present paper was well under way when the Barzel results were pointed out to me by Charles Knoeber. A paper by Manchester (1976) also presents (in an appendix) a regression related to those estimated below. However, Manchester does not interpret his tax rate coefficients in that paper.

[3] For multiproduct firms, optimal pricing decisions require attention to cross elasticities of demand. I will simply ignore this refinement and implicitly assume that it is not empirically important. In what follows I use the term "firm" to mean the relevant pricing decision maker for the given product.

is captured by state dummy variables and does not bias the coefficient of R_{it}.

IV. Empirical Results and Interpretations

The major empirical results of this study are reported in table 1. In this section, these results are presented and interpreted. A potential alternative interpretation suggested by Barzel (1976) is also considered.

The estimates reported are based on annual tax rates and price data for the years 1954–78 for 47 states and the District of Columbia. Alaska, Hawaii, and New Hampshire are excluded from the sample for the analysis reported in the table because of lack of data and peculiarities in their tax laws. (See n. 11 below for more discussion of the New Hampshire case.)

Information on cigarette prices and tax rates is collected and published by the Tobacco Tax Council. The "weighted-average price per package" published for each state and year uses national weights for type of cigarette (regular, king, 100 mm) and for type of transaction (carton, single pack, machine). Wholesalers around the country are asked to provide estimates of the average retail price in their area for each type of cigarette and type of sale. These numbers are turned into state averages by the council. (This information is from correspondence with the very helpful personnel at the Tobacco Tax Council.)

Table 1 shows estimates of the linear model of equation (4) under three alternative specifications of the X_{it} vector and the disturbance term. The table reports the estimated θ and, to aid in interpretation, the implied price elasticity, η. In row 1, X_{it} consists of a dummy variable for each state and for each year. The tax rate effect in this row is 1.074, which says, for example, that a 10 cent difference in tax rates causes a 10.74 cent increase in price. The coefficient is significantly above 1.0 (standard error = 0.013), which is support for rejecting the hypothesis of atomistic competition in this industry. However, the implied price elasticity is −13.5, which indicates a rather flat demand curve. Row 2 replaces year dummies with the National Consumer Price Index and drops the state dummy variables, while row 3 shows estimates from an application of a variance-components model with the price index as the only additional explanatory variable. The estimates of θ for these rows are 1.029 and 1.069, and both are significantly above 1.0. Other specifications of X_{it} and U_{it} yielded estimates of θ similar to those reported.

This basic model was also estimated separately for each year using states as the unit of observation. The estimates of θ were generally in

TABLE 1

ESTIMATES OF THE EFFECT OF THE TAX RATE ON CIGARETTE PRICE IN ALTERNATIVE SPECIFICATIONS OF THE LINEAR MODEL (Eq. [4])

Estimated θ	Implied Price Elasticity η	Other Variables Included*	Method
1.074 (.013)	−13.5	DY_t, DS_i	OLS
1.029 (.011)	−34.5	NCPI	OLS
1.069 (.013)	−14.5	NCPI	VC†

NOTE.—The R^2's in each of these models were high, those with dummy variables were about .99, and those without the dummies were around .95. There were 1,200 observations (25 years for each of 48 states). SEs in parentheses.

* DY_t is a dummy variable for each year with 1978 excluded, DS_i is a dummy variable for each state with Wyoming excluded, and NCPI is the national consumer price index for each year.

† In this equation a variance-component (VC) error scheme with an individual random effect for each state and each year was used. See Fuller and Battese (1974) for a discussion.

the range reported in table 1 for the pooled model and were not significantly different from one another. No trend in these estimates was noted. Time-series changes in the industry do not seem to have caused biases in the estimates in table 1.[9]

Two further comments should be made in relation to these results. First, these estimates are *not* consistent with an effective cartel in the cigarette industry. The firm-level elasticity estimates are well below the range of the industry demand elasticity estimates usually found for cigarettes. (These are generally between −0.3 and −0.8.)[10] Second, the effect of monopoly power as a source of price variation in the industry is small compared with tax rate differences and other cost differences over time and space.

[9] In fitting the nonlinear specifications presented in n. 7 above, the hypothesis that $\lambda = 1$ is not rejected. For (4′) $\hat{\lambda} = 0.972$ with SE = 0.027, and for (4″) $\hat{\lambda} = 0.981$ with SE = 0.026. These tests support the appropriateness of the linear model results presented in table 1. The estimate of γ from (4′) is 1.029, and from (4″) it is 1.062. These, together with the λ estimates above, imply values of θ in the range of those in table 1 or slightly lower. This supports the hypothesis that while perfectly competitive pricing behavior is rejected, monopoly power is not a major factor in determining cigarette price differentials. The hypothesis that interstate smuggling would imply that price is a concave function of tax rate, as discussed in n. 6 above, is not supported by these results. As a further nonlinear specification, a squared tax rate term was added to the separate yearly price equations. In most years the $(R_{it})^2$ variable had an insignificant effect on price. In only a few years out of 25 were the coefficients for R_{it} positive and $(R_{it})^2$ negative.

[10] In related work (Sumner 1979), using cross-section time-series data for the last 10 years, which attempts to indirectly quantify the extent of interstate cigarette bootlegging ("buttlegging"), I find the national *industry-level* elasticity of demand for cigarettes to be approximately −0.40. This is in the range of other estimates from time-series data.

Barzel (1976) has suggested an alternative interpretation for the coefficient on R_{it} in the pricing equation. He argues that a higher per unit tax on a good will induce a higher quality per unit to be sold. Since the relative price of quality is lower in high-tax areas, substitution of quality for quantity will occur. The after-tax price per unit will be higher in a higher tax rate area to reflect the higher quality of the product. He also notes that ad valorem taxes fixed as a percentage of retail price (as was the practice in New Hampshire prior to 1975) would not have this effect.

As Barzel points out, products actually traded are combinations of many characteristics, most of which are extremely difficult for either the government or the empirical economist to measure. Thus, when Barzel argues that the quality of cigarettes varies positively with the tax, no direct measurement of quality changes could reject that hypothesis. Barzel turns instead to price data on cigarettes across years and states as published by the Tobacco Tax Council. The nature of these data is crucial. As noted above, the price data available for cigarettes have the most obvious quality variations (type of sale and type of cigarette) held fixed across states. Furthermore, given centralized manufacturing and national distribution of brands, there is little scope for direct product-quality variation. This leads one to question the strength of the expected Barzel effect in these data even if it were present in the real world. Hence it would appear that the "elasticity of demand" interpretation of θ is a serious contender with Barzel's quality shifter for explaining a θ coefficient greater than 1.0. Finally, note that the Barzel effect would cause an upward bias in the measurement of monopoly power so that there is an even stronger case for concluding that the cigarette industry does not act as a cartel.[11]

[11] As further consideration of the strength of the Barzel effect in cigarettes, some estimation was done which included data from New Hampshire with the other states. These results were essentially identical to those reported in the table. New Hampshire, which had an ad valorem cigarette tax until 1975, was found to have a strongly negative state dummy variable. However, it may be the case that New Hampshire has a low price of cigarettes net of taxes not because its tax was ad valorem as suggested by Barzel but for the same reason that the per capita sales of cigarettes in New Hampshire are more than double the average of the other New England states. That is, residents of nearby states buy a significant portion of their cigarettes in New Hampshire. Given size economies for sales transactions, a high proportion of this interstate trade would tend to reduce the average price net of taxes in New Hampshire. Barzel's ad valorem tax hypothesis implies that the special position of New Hampshire would have evaporated after 1975 (when a fixed per unit tax replaced the ad valorem tax). The transaction cost hypothesis (among others) predicts that there will be no effect of the 1975 law change. I found that the New Hampshire residuals from the linear model estimated for each year are just as negative after 1975 as before 1975, which indicates that the ad valorem tax was not the different New Hampshire characteristic. This is not to suggest that the Barzel effect is not theoretically sound but, rather, that interstate cigarette prices may not be useful to test the theory.

The statistical results presented in table 1 support rejection of the hypothesis of atomistic competition in the cigarette industry. The demand curves facing cigarette firms are not flat but are close to it by most standards. Taking $\theta = 1.05$, a 1 percent rise in the price of a firm's product would cause a 20 percent fall in the quantity sold. If the cigarette prices are above marginal cost by 5 percent, monopoly power adds approximately 3 cents to the average price of a package of cigarettes.

This paper has proposed and demonstrated the applicability of a scheme for measuring monopoly behavior that may be relevant to several different industries. For example, application to the U.S. gasoline market is currently being pursued. Data requirements are information on product price and some factor that has a directly measurable per unit cost. Variation over time or space or some other dimension is required for statistical implementation.

References

Bain, Joe S. *Industrial Organization*. 2d ed. New York: Wiley, 1968.

Barzel, Yoram. "An Alternative Approach to the Analysis of Taxation." *J.P.E.* 84, no. 6 (December 1976): 1177–97.

Friedman, Milton. "The Methodology of Positive Economics." In *Essays in Positive Economics* Chicago: Univ. Chicago Press, 1953.

Fuller, Wayne A., and Battese, George E. "Estimation of Linear Models with Cross-Error Structure." *J. Econometrics* 2, no. 1 (May 1974): 67–78.

Johnson, Terry R. "Additional Evidence on the Effects of Alternative Taxes on Cigarette Prices." *J.P.E.* 86, no. 1, pt. 1 (April 1978): 325–28

Manchester, Paul B. "Interstate Cigarette Smuggling." *Public Finance Q.* 4, no. 2 (April 1976): 225–38.

National Association of Tobacco Distributors. *Coordinator*. New York: NATD, 1979.

Nicholls, William H. "The Tobacco Case of 1946." *A.E.R. Papers and Proc.* 39, no. 3 (May 1949): 284–96.

Stigler, George J. *The Theory of Price*. 3d ed. New York: Macmillan, 1966.

Sumner, Daniel A. "Cigarette Smuggling and Estimates of the Demand to Consume and the Demand to Smuggle." Paper presented at a workshop on agricultural economics at North Carolina State Univ., Raleigh, November 1979.

Tennant, Richard B. "The Cigarette Industry." In *The Structure of American Industry*, edited by Walter Adams. 4th ed. New York: Macmillan, 1971.

Swedish Tax Rates, Labor Supply, and Tax Revenues

Charles E. Stuart

University of California, Santa Barbara and University of Lund

Effective marginal tax rates on labor income for the "representative" Swede have increased from roughly 50 percent in 1959 to 80 percent today. The effects of this increase in the level of taxation are examined using a two-sector model parameterized to correspond to the Swedish economy. The model contains a single household which allocates labor to either taxed (essentially market) or untaxed (largely household) uses. The estimated long-run effects are sufficient to explain up to 75 percent of the recent decline in the measured growth rate of the Swedish GNP. Calculations of total tax revenues are also derived from the model. These peak when the tax rate is approximately 70 percent, indicating that Sweden is presently on the downward-sloping portion of its "Laffer Curve."

This paper considers the statement that high taxes can have serious consequences. Given the importance of the issue and the attention paid it in informal discussions, it is surprising that little formal analysis has been directed at it. The present treatment focuses on the Swedish experience. This is a case of current events steering an economist's eye. The growth of the Swedish GNP has slowed considerably in recent years, falling from 4.4 percent per year during 1962–67, to 2.9 percent during 1967–72, to 1.4 percent during 1972–77. At the same time, tax revenues as a portion of GNP have

Thanks are due to Syed Ahsan, Michael Bordo, Geoff Carliner, Ingemar Hansson, Christina Jonung, Bjorn Lindgren, Goran Normann, Goran Skogh, and an anonymous referee. Useful comments were provided by participants of seminars which were held at Carleton University, Ottawa, Concordia College, Montreal, and in Lund and Uppsala. The research support of the Bank of Sweden Tercentenary Foundation is gratefully acknowledged.

[*Journal of Political Economy*, 1981, vol. 89, no. 5]
© 1981 by The University of Chicago. 0022-3808/81/8905-0012$01.50

increased from 0.388 in 1962 to 0.627 in 1977.[1] While the exercise of presenting these numbers proves nothing, it does indicate that a study of the allocational effects of high taxes may provide insight into the faltering of the Swedish economy.

There are many possible explanations of the Swedish economic stagnation, and the goal here is not to test competing explanations against each other. Instead, the strategy is to concentrate on household willingness to supply labor to a taxed (market) sector and to ask whether changes in the tax rate might be expected to have significant long-run effects on the equilibrium quantity of labor supplied. The first major conclusion of this paper is that the calculated "supply-side" effects of high taxes are of sufficient magnitude to at least partially explain the observed stagnation.

A useful summary statistic for the effect of the level of taxation on labor supply is the (marginal) tax rate which maximizes total tax revenues (or, in more popular terms, the rate which yields a maximum of the "Laffer Curve" for labor). The usefulness of this rate—hereafter denoted t^*—is that as taxes increase and t^* is approached, the marginal social cost of additional tax revenues becomes infinite. While the socially optimal tax rate cannot be determined without regard to a welfare criterion, rather extreme assumptions must be made about the nature of social welfare in order to imply an optimal tax rate greater than t^*, since reductions in the tax rate from levels exceeding t^* would increase both the worth of output and tax revenues.[2] Values of t^* are calculated here; depending on the assumptions used, the best estimates of t^* lie in the interval 69–73 percent. This leads to the second major conclusion of this paper, that actual tax rates in Sweden are in excess of t^*. The aggregate marginal tax rate (defined below) in Sweden reached the 70 percent level in the early 1970s and is slightly in excess of 80 percent today.

The vehicle of analysis is a two-sector (Harberger) model where an aggregate household allocates labor to either a taxed or an untaxed sector.[3] The former corresponds closely to the legal, market portion of the economy and is typically represented in national income accounts. The latter includes home production and activities yielding undeclared income (illicit tax evasion); leisure may also be included. The advantage of such a general equilibrium framework is that it

[1] These include direct and indirect taxes, social insurance premiums, and estate and gift taxes.

[2] See Cooter and Helpman (1974) for a discussion of optimal taxation under different welfare functions.

[3] A similar sectorial division was used by Boskin (1975). As in Harberger's (1962) treatment, Boskin's analysis is not based on an explicit utility function; demand elasticities are instead taken as parameters. The present model does explicitly specify a utility function; in this respect, it is similar to the model of Shoven and Whalley (1972).

yields a clear picture of how taxation affects the relative benefits of different uses of labor and how this in turn impinges on labor supply and output. An alternative and simpler approach would be to interpret econometric estimates of labor-supply elasticities as reflecting the net-wage sensitivity of labor supplied to taxed uses and to assume that any tax increase leads to an exactly corresponding reduction in the net wage of taxed labor. However, this alternative approach suffers from several weaknesses: (1) it is partial; (2) it would in practice be based on constant elasticity estimates, whereas (as will be seen below) labor-supply elasticities should vary with the marginal tax rate; (3) it does not yield estimates of output under different tax rates and thus makes simple welfare comparisons difficult; and (4) indeed, the set of activities measured in traditional labor-supply studies does not coincide exactly with the set of taxed uses of labor. Despite this care in the choice of a general equilibrium framework, the present study should still be regarded as more of a first analysis than a final answer to the problem of the supply-side effects of taxation. Perhaps the two most severe limitations of the analysis are the highly aggregated treatment of untaxed uses of labor and the relative lack of attention given capital accumulation.

In the next section the model is formulated and parameterized. Results are presented in Section II. Section III is concerned with the predictive power of the results. Concluding remarks are made in Section IV.

I. The Model

A. Structure

Consider a highly aggregated economy with two sectors, a taxed sector, in which employed labor is subject to a tax, and an untaxed sector, where supplied labor is not taxed. One may think of the untaxed sector as involving not only illicit tax evasion via undeclared exchange but also fully legal forms of tax evasion, such as painting one's own home instead of hiring a professional painter, cooking fancy meals instead of eating out at a fine restaurant, repairing an automobile oneself, sewing clothes, watching children, cleaning the house instead of hiring domestic help, etc. It is also possible to interpret leisure as an output of the untaxed sector.[4] The untaxed sector thus consists largely of home production activities. On the other hand, the taxed sector is that part of the economy which is traditionally treated in national income accounts.

[4] A more detailed model of home production, leisure, and market work is contained in Gronau (1977).

Each of these sectors is represented by a Cobb-Douglas production function, a choice motivated by the observed statistical stability of these functions (see Douglas 1976). Letting Y_1 and Y_2 denote the outputs of the taxed and untaxed sectors, respectively, and letting L_1 and L_2 signify the corresponding uses of labor in these sectors, the production functions are:

$$Y_1 = AL_1^a, \tag{1a}$$

$$Y_2 = BL_2^b, \tag{1b}$$

where A, a, B, and b are parameters. Formally, capital is subsumed in the coefficients A and B, while a and b represent labor's shares in the two sectors. This formulation implicitly assumes perfectly inelastic supplies of capital to both sectors so that capital remains constant during the simulations. Aside from simplifying the analysis, such a treatment avoids the difficult empirical problem of estimating the capital stock in the untaxed sector. In addition, relaxing the constant capital assumption—by letting the supply of capital to either sector depend positively on sectorial net return—would cause the calculated value of t^* to fall and hence strengthen the conclusion that the aggregate marginal tax rate in Sweden today exceeds t^*. This is because an increase in the tax rate on labor income would then give rise to the following additional negative feedback on the tax base: As the amount of taxed-sector labor fell, capital's marginal product in taxed production would also fall, decreasing the demand and hence equilibrium quantity of taxed-sector capital, thereby reducing the gross wage obtained in the taxed sector. The result is that the taxed sector's wage bill would be lower (than it would have been with fixed capital).[5]

Household choice dictates the allocation of labor to the taxed and untaxed sectors; the household supply constraint for labor is

$$L_1 + L_2 = L. \tag{2}$$

For simplicity, all taxes are treated as impinging on households. The problem of the household—and again the analysis is simplified by considering the behavior of a single, aggregate household—is to

[5] An additional source of erosion in the tax base for labor income can occur if capital is variable and if the tax under consideration falls on capital as well as labor income. This occurs with income taxes. Here, an increase in the tax rate will cause an increase (or no decrease) in the purchase price of capital, and less capital will be used in the taxed sector. This in turn reduces the taxed sector's wage bill. The net result is that tax revenues must peak at a lower tax rate when capital is taxed than when it is untaxed. Again, the assumption of fixed capital leads to the "correct" bias, relaxing it while simultaneously allowing taxation of capital can only strengthen the conclusion that the aggregate marginal tax rate in Sweden today exceeds t^*.

maximize a utility function dependent on Y_1 and Y_2. Two alternative forms for utility are proposed:

$$U = Y_1 + Y_2, \tag{3a}$$

$$U = Y_1^{\alpha} Y_2^{1-\alpha}. \tag{3b}$$

The former encompasses the hypothesis that the household is an income maximizer and that Y_1 and Y_2 are perfect substitutes. The latter formulation allows imperfect substitutability between Y_1 and Y_2.

From the point of view of the household, tax revenues are redistributed on a lump-sum basis, and capital incomes are obtained on the same basis. The amount of Y_1 consumed by the household is thus

$$Y_1 = (1 - t)wL_1 + R + I, \tag{4}$$

where t is the average tax rate, w is the wage, R is redistributions, I is capital income, and where the output of the taxed sector is taken as numeraire. It is assumed that the government budget balances in equilibrium, $R = twL_1$, and that capital's share of Y_1 is $1 - a$, so $I = (1 - a)Y_1$. These conditions guarantee that the amounts of Y_1 produced and consumed are equal. The amount of Y_2 consumed is taken as the amount produced, as no taxes are levied on Y_2 or L_2.

The first-order condition for a maximization of utility subject to (1b), (2), and (4) is

$$(1 - t')w = \left(\frac{\partial U}{\partial Y_2} \Big/ \frac{\partial U}{\partial Y_1}\right)(bBL_2^{b-1}), \tag{5}$$

where the left-hand side is marginal retentions of Y_1 from a unit of labor allocated to taxed production, given marginal tax rate t', and where the two terms in parentheses on the right-hand side are the marginal rate of substitution and the marginal product of labor in untaxed production, respectively. The marginal rate of substitution is either one, for (3a), or $[(1 - \alpha)/\alpha] \cdot (Y_1/Y_2)$, for (3b). The import of (5) is that taxes affect the household by altering the relative marginal benefits of labor allocated to the two sectors.

The model is closed by assuming that the wage in the taxed sector is determined under competitive conditions:

$$w = aAL_1^{a-1}. \tag{6}$$

B. *Parameterization*

To use the framework presented above as a basis for simulations, it is necessary to obtain parameter estimates of A, a, B, b, and α, as well as to estimate the amount of labor available, L. A summary of the method of parameterization is presented here; a more detailed statistical discussion is contained in the Appendix.

The technique for parameterizing the model is to pick an appropriate base year, choose units of measurement for labor so that the wage equals 1 during that year, and then to find parameters consistent with actual economic behavior in Sweden for that year. The choice of the base year for the study was 1969 as it was a year of "normal" business activity; that is, it was neither a peak nor a trough of a business cycle (1968 was a low, while a peak occurred in 1970–71).

First, a is taken as labor's share of measured GNP (see Appendix, item 1). This yields $a = 0.752$, whence a knowledge of Y_1, L_1, and a allows $A = 4.19$ to be calculated in accordance with (1a).

Estimating the shape of the untaxed sector's production function is more difficult. Due to the long-term statistical stability of labor's share in Cobb-Douglas functions and the consideration that many of the commodities produced by one sector can also be produced by the other, the naive assumption that $b = a = 0.752$ is used for most simulations. However, given the desirability of conducting sensitivity analysis, as well as the fact that 0.752 is somewhat greater than the values of labor's share found by many other authors (for discussion, see Douglas [1976]), the alternative assumption that $b = 0.5$ is applied to two sets of simulations.

Estimates of the amount of time spent in untaxed production and the value of labor in untaxed production are derived from time-budget and wage data, with the latter calculation based on a variant of the method used by Nordhaus and Tobin (1972) in their work on "measures of economic welfare" ("MEW"s) (see Appendix, item 2). In most of the simulations, L_2 and Y_2 are defined as not including leisure; this implies $L_2 = 102.63$ and a value of untaxed labor of 54.45.[6] In the remaining cases leisure is included, yielding $L_2 = 205.26$ and a value of untaxed labor of 105.14 for 1969. Aside from the obvious benefit of including leisure as an alternative use of household time, the exercise of running simulations on L_2 defined to include leisure has the advantage of providing a sensitivity test on the otherwise fairly coarse estimates of untaxed production. Such a sensitivity test is particularly valuable, since estimated untaxed produc-

[6] Hawrylyshyn (1976) surveys earlier estimates of the value of home production and discusses the appropriateness of the methods used. He also recalculates earlier estimates to put them on a comparable and more correct basis. He concludes that the value of labor in home production is 35–40 percent of GNP (see also Murphy [1978] and Adler and Hawrylyshyn [1978] for more recent estimates for the United States and Canada). The ratio of value of labor in home production to GNP in the present study is 40 percent. One might have expected a higher value for Sweden in 1969, since greater taxes should imply relatively more time devoted to untaxed (home) production. Note, however, that these earlier studies use a 20 percent tax rate in deriving after-tax wages; the rates used in the present study are 30–40 percent

tion is calculated as home production and thus fails to capture pr duction which is untaxed due to illegal tax evasion.[7]

A knowledge of the amount of time spent in untaxed productic and of the value of labor in untaxed production is then used to find value of B for each case. The method is to solve

$$b = \frac{\text{(value of labor in untaxed production)}}{Y_2} \qquad (7$$

for Y_2 and then to substitute into (1b).

Throughout most of the simulations, the total amount of lab available to the household is assumed to be perfectly inelastic. Th value of L is thus the sum of L_1 and L_2 for 1969. This procedure migl be objected to since greater after-tax wages may create an inducemer for households to supply more labor. To evaluate the sensitivity of th results to this sort of effect (which might also arise if net immigratio depends on after-tax wages) as well as to treat an alternative, "ind rect" method of introducing leisure, several simulations are cor ducted using the specification

$$L = c[w(1 - t')]^{0.1}, \qquad (8$$

where c is chosen so that $L = 205.26$ when the 1969 values of w and t are inserted. This formulation assumes a constant, compensated elas ticity of supply of labor of 0.1.[8]

Finally, data on taxes are needed to complete the parameterizatior To get a feel for how marginal taxes should be interpreted in th model, it is useful to think for the moment of the taxed and untaxe sectors as being market and nonmarket sectors, respectively, and t ask, What taxes will an individual face if he operates in the marke sector? First, the individual sells his labor (L_1). The employer pays wage w. Before the employee receives his paycheck, payroll taxes ar drawn. Next, income taxes are paid and the individual is left with sum of money. This money is taken to a commodity market to b reconverted into goods and services. Indirect taxes are then drawn Thus the tax rate relevant to the model includes payroll, income, an

[7] Unfortunately, no careful economy-wide estimates of the extent of illegal ta evasion exist for Sweden For the United States, recent work by Feige (1979, 198 suggests that the size of the "monetary unobserved sector" (i.e., the volume of unre ported transactions which use money as a medium of exchange) has increased greatly i recent years. Feige calculates that the size of this sector was roughly 5 percent o measured GNP in the early 1960s but amounts to 25–30 percent of measured GN today, with most of the growth taking place in the 1970s. (In correspondence, Feig expresses greater confidence in his estimates of growth rates for this sector than in th estimated absolute sizes of the sector.)

[8] Actually the elasticity is almost compensated since (1) all tax revenues are returne to the household on a lump-sum basis in the model, but (2) marginal increases in taxe may decrease national income slightly.

TABLE 1

SWEDISH TAX RATES 1954–74

Year	Average Tax (t)	Marginal Tax (t')
1954	.271	.419
1959	.321	491
1964	.393	583
1969	.468	.655
1974	.557	.764

indirect taxes. In addition, certain government transfers are income-indexed, so a correction must be made for the marginal loss of subsidies due to increased labor (L_1) income. The starting point for a calculation of these "total" tax rates is a study by Jakobsson and Normann (1974), where the aggregate average and marginal income tax rates for the Swedish economy were estimated.[9] Table 1 contains calculated values of the taxes used in the model (for details, see Appendix, item 3).

On the basis of these rates, it is assumed that increases in tax rates satisfy

$$t = 0.829t' - 0.075. \tag{9}$$

The slope of this function is calculated as the change in marginal tax rates divided by the change in average tax rates from 1954 to 1974. The constant (−0.075) adjusts so that (9) holds exactly for 1969.

Four scenarios (sets of simulations) are considered under the income-maximization hypothesis. Scenario A is the basic simulation; it has $b = 0.752$, leisure not included in L_2, and constant L. In scenario B, the value of b is reduced from 0.752 to 0.5 as a sensitivity test. Scenarios A and C are identical except that leisure is included in L_2 in the latter. Finally, scenario D differs from scenario A only in that L is assumed to vary according to (8).

It is interesting that if the parameter values for scenarios A–D and 1969 values of L_1 and L_2 are inserted into the first-order condition for a maximization of income (eq. [5] with a unitary MRS), it is found that the implied values of the marginal tax rate are significantly less than the rate which actually prevailed in 1969. These implied rates are between 0.47 and 0.49, which approximate the marginal tax rates in 1959–60. The interpretation is that if the income-maximization hy-

[9] This was done by formulating an aggregate total tax function based on tax law and population (mainly income-distribution) data, simulating different total income–total tax values, and regressing total taxes on total income. Rates for 1974 were supplied by G. Normann.

TABLE 2

SCENARIOS AND PARAMETER VALUES

Scenario	b	B	α	L	Explanation
			Income Maximization		
A.	752	2 22	...	205 26	Basic
B	5	10.75	.	205 26	Sensitivity analysis, b
C	.752	2 55	...	307.89	Leisure included in L_2
D	.752	2 22	..	$220[w(1 - t')]^{0.1}$	L depends on after-tax wage
			Cobb-Douglas Utility Maximization		
A'.	752	2 22	.744	205 26	Basic
B'.	5	10.75	.659	205.26	Sensitivity analysis, b
C'	.752	2 55	592	307.89	Leisure included in L_2
D'	752	2 22	.744	$228[w(1 - t')]^{0.1}$	L depends on after-tax wage
E'.	.752	2.22	.706	205 26	5-year adjustment lag

NOTE — It is assumed that a = 0 752 and A = 4 19 throughout

pothesis is correct, there was a 9–10-year time lag in the adjustment of household behavior to tax rates. Such a long time lag may seem unreasonably high but is not impossible. Attitudes toward labor market participation and the "correct" amount of work are likely to change slowly, especially given the rigidity of social institutions such as the 8-hour day, the 40-hour week, etc.

When the household is assumed to maximize the Cobb-Douglas utility function (3b), a parameter value for α must be specified. This value was chosen so as to be consistent with (or the root of) the first-order condition (5) for 1969 when all other parameters, endogenous variables, and the tax rate were inserted. Five scenarios are investigated for this utility function. The first four are similar to the four scenarios considered for income maximization. In the fifth case, it is assumed that a 5-year lag existed in the adjustment of L_1 and L_2 to changes in the tax rate. This is accomplished by letting the tax rate at which the model is parameterized be the marginal rate which occurred in 1964 (which, it may be noted, occupied a position in its business cycle similar to that of 1969).

A summary of the scenarios is given in table 2.

II. Results

Within the framework presented above, simulations are conducted by plugging various marginal tax rates into condition (5) and then solving numerically for the values of L_1 and L_2 and, hence, also of Y_1 and Y_2, which solve (2), (5), (6), and, when applicable, (8). Equation (9) is

TABLE 3

RESULTS

Scenario and Marginal Tax (%)	$\%\Delta L_1$	$\%\Delta Y_1$	Y_2/Y_1	t^* (%)
A:				
80	−22.3	−15.6	10.2	
65	−10.0	−7.41	1.86	46
50	−4.63	−3.44	.63	
B				
80	−19.0	−14.6	7.92	
65	−7.41	−5.64	1.81	51
50	−3.12	−2.29	.84	
C:				
80	−19.4	−15.3	17.7	
65	−10.57	−7.91	3.26	43
50	−5.69	−4.26	1.10	
D:				
80	−19.8	−15.7	10.2	
65	−9.78	−7.45	1.86	46
50	−4.73	−3.51	.63	
A′:				
80	−3.24	−2.25	.80	
65	−1.46	−1.10	.52	73
50	−.83	−.65	.40	
B′				
80	−3.23	2.43	1.13	
65	−1.36	−1.10	.79	73
50	−.83	.58	.64	
C′				
80	3.98	−3.00	1.54	
65	−1.95	−1.39	1.01	69
50	−1.24	−.87	.78	
D′:				
80	3.68	−2.72	.80	
65	−1.75	−1.25	.52	71
50	−.96	−.76	.40	
E′				
80	−3.51	−2.68	.92	
65	−1.61	−1.19	.60	72
50	−.99	−.69	.46	

NOTE.—Percentage changes are calculated using differences in natural logarithms for tax rate increases of 1 percent (e.g., from 80 percent to 81 percent).

then used to determine total tax revenues (i.e., to derive the "Laffer Curve").

The results of the simulations are summarized in table 3. Since the model is parameterized using data from 1969, when the marginal tax

rate was 65.5 percent, simulations at tax rates significantly greater or less than 65.5 percent may be less reliable than simulations at rates close to 65.5 percent. For this reason, results are reported for the 65.5 percent marginal tax rate, as well as for ±15 percent changes in marginal rates. Reported in the columns marked "$\%\Delta L_1$" and "$\%\Delta Y_1$" are the percentage changes in L_1 and Y_1 induced by a 1 percent increase in marginal taxes (i.e., from 80 percent to 81 percent, 65 percent to 66 percent, and 50 percent to 51 percent). The column Y_2/Y_1 gives the ratio of outputs of the two sectors. Finally, the tax rate which maximizes total tax revenues, t^*, is presented.

Several facets of the results stand out. First, the scenarios based on income maximization yield quite striking results. As tax rates increase from the 50 percent to 80 percent brackets, the ratio of untaxed to taxed output rises dramatically, as do the sensitivities of taxed-sector employment and output to small tax increases. The scenarios based on Cobb-Douglas utility maximization give somewhat less striking results, but the general pattern is the same. For instance, a shift from 50 percent to 80 percent taxes generally implies a doubling in the size of untaxed production relative to taxed production.

An important aspect of the results is that, as tax rates rise in 1 percent increments from 50 percent to 80 percent, the effects of successive increments on taxed-sector employment and output become increasingly larger. Part of the explanation is that the optimizing household is interested in the rate of marginal retentions (1 minus the marginal tax rate), and marginal retentions fall by increasing percentages as marginal taxes rise incrementally. For example, if equations (5) and (6) are consulted and the marginal rate of substitution is assumed constant (the income-maximization hypothesis), then a 1 percent tax increase around the 50 percent bracket requires a change in L_1 and L_2 such that the ratio of labor's marginal products in the taxed and untaxed sectors rises by 2 percent, but at an 80 percent tax rate, the same 1 percent tax rise requires a 5 percent increase in the ratio of labor's marginal products.

The calculated values of t^* are on the order of 50 percent under income maximization and 70 percent under Cobb-Douglas utility maximization. The sensitivity of t^* around the 50 percent and 70 percent figures is small, indicating a fair degree of robustness to the details of the parameterization of the model. The robustness of the calculated values of t^* is enhanced by the fact that these values lie relatively near the tax rate which prevailed in 1969, the year used for parameterization. Further, the inclusion of leisure into L_2 leads to relatively small declines in t^* of 3–4 percent from the basic scenarios (A and A′).

The values of t^* calculated here are for taxes on labor income alone.

However, there are two reasons why increases in the aggregate tax rate used in the present analysis might also affect the extent of taxation of capital: (1) except for the payroll tax, all of the components of this tax rate also apply, at least partially, to income from capital; and (2) in attempting to finance a greater volume of public spending, it is likely that the government would utilize (or has utilized) a mix of tax increases for both labor and capital incomes instead of relying solely on increased taxation of labor. One might therefore ask, How would the calculated values of t^* change if t^* were redefined to imply a maximum of tax revenues from both labor and capital incomes?

To investigate this issue, it is useful to retain for the moment the assumption that capital is fixed. In this context, a simultaneous increase in the tax rates on labor and capital will have a positive initial impact on revenues from capital taxation but a negative feedback effect since national income, and hence the amount of capital income subject to taxation, will decline. The following calculations indicate that the net effect on the previously calculated values of t^* is likely to be negligible. In 1969, returns to capital were 33.8 (billions of kronor). Total taxes collected in the economy were 62.8, of which 48.0 were levied on labor income. Of the remainder, something like 14.0 can be regarded as taxes on capital, whence the average tax rate on capital was 0.41 (= 14/33.8). From 1954 to 1974, the average tax rate on labor income rose by 0.286 (see table 1). This rise was partially due to a hike in the average income tax rate of 0.15 and to a rise of 0.02 in the indirect tax component of the rate on labor income. One possible set of assumptions, then, is that the average tax rate on capital rose by 0.17 and that each 1 percent increase in the average tax on labor thus "caused" a 0.6 percent (= 0.17/0.286) increase in the average tax on capital, so that the tax on capital satisfied $t_K = 0.6t_L + 0.13$ (with 0.13 chosen so the 1969 values, $t_K = 0.41$ and $t_L = 0.47$, hold simultaneously). Then, with capital in perfectly inelastic supply and capital's share of Y_1 equal to 0.248 (= 1 − 0.752):

$$\text{Total tax revenues} = \text{taxes on labor} + (0.6t_L + .13)(0.248Y_1).$$

Values for taxes on labor, t_L, and Y_1 can be obtained from the simulations. It is then a simple matter to calculate the modified tax rates which maximize total tax revenues, including taxes on capital. For scenarios A–D and A′–E′, these are, respectively, 0.45, 0.50, 0.41, 0.44, 0.72, 0.72, 0.68, 0.70, and 0.71. Correcting for taxes on capital in this way thus reduces the table 3 estimates of t^* by about 0.01.

Of course, it might be argued that the relationship between tax rates assumed in this formulation ($t_K = 0.6t_L + 0.13$) understates the tendency for taxes on capital to rise where the rate on labor income

goes up, and that this yields an overly pessimistic view of the ability of the government to increase tax revenues. Perhaps a more plausible formulation would be that t_L and t_K lie in constant proportion to each other, $t_L = \beta t_K$ (and, say, $\beta = 0.47/0.41$). In this case, observe that:

$$\text{Total tax revenues} = t_L a Y_1 + t_K(1 - a)Y_1 = t_L Y_1\left(a + \frac{1 - a}{\beta}\right),$$

and a maximum of $t_L Y_1$ must imply simultaneous maxima for both taxes on labor ($t_L a Y_1$) and for total tax revenues.

Allowing in this way for capital taxation but keeping capital fixed thus has little or no effect on the calculated values of t^*. What then is the effect of relaxing the assumption that capital is fixed? There are two reasons why this must cause the calculated values of t^* to fall. First, if the total supply of capital to the economy depends positively on capital's net return—as the work of Boskin (1978) suggests—then increased taxes on capital will reduce the amount of capital in the economy, causing a decline in income and resultant declines in tax revenues from both capital and labor incomes. Second, to the extent that the taxed sectors for labor and capital coincide, a rise in the rate on capital will drive capital out of the taxed sector, and again taxed sector income, and hence tax revenues from both capital and labor, will fall.

III. Does the Model Predict Well?

It turns out that the simulations based on Cobb-Douglas utility maximization predict fairly well the actual behavior of output and employment during the period of rising taxes in Sweden. The income maximization hypothesis, on the other hand, leads to predicted declines in output and employment which greatly exceed actual declines. This allows a tentative rejection of the income maximization hypothesis. In what follows, rough calculations of the effects of recent increases in marginal taxes in Sweden are made under the assumption of Cobb-Douglas utility maximization. These predictions are then compared with the actual development of output and employment.

First, it was noted at the outset that yearly growth rates declined from 4.4 percent in the early and mid-1960s to 1.4 percent in the mid- and late 1970s. This amounts to a decline in the growth rate of 3 percent per year. Over this period, marginal tax rates rose an average of something like 1.5 percent per year, increasing from around 55 percent to upward of 80 percent. In 1972, the marginal tax rate was 70–71 percent. A calculation of $\%\Delta Y_1$ at a 70 percent rate yields values of between 1.4 percent and 1.7 percent, depending on the scenario chosen. Applying a figure of 1.5 percent to the entire period

1972–77 (a procedure which effectively assumes a slight time lag in the adjustment of L_1 to changes in tax rates) implies a 2.25 percent output decline per year (i.e., the 1.5 percent yearly increase in tax rates × the 1.5 percent decline in output per percent increase in taxes). This calculated yearly output decline due to rising taxes amounts to 75 percent of the 3 percent fall in the growth rate.[10]

The model also predicts a decline in employment. During the period 1964–69, marginal taxes rose from 58.3 percent to 65.5 percent, which implies average yearly reductions in L_1 of between 1.8 percent and 2.5 percent, depending on the scenario, if there was no adjustment lag. From 1964 to 1969, the number of hours worked divided by the population of working age (16–74) actually fell at a 1.6 percent yearly rate. This is a fairly good fit. Moreover, a legislated reduction in the workweek was effected in 1970–72, and this may be interpreted as partially representing an adjustment to the tax rates in effect in the late 1960s.

During the period 1969–77, the ratio of hours worked divided by the population of working age fell roughly 7 percent. The increase in marginal taxes which occurred during this period should, according to the simulations, have led to reductions in employment significantly in excess of this figure. While there is no reason to believe that the economy was particularly lax in adjusting to its equilibrium level of employment in 1969 (indeed, there was net immigration throughout the late 1960s), there are two reasons for supposing such a lag existed in the 1970s. First, the very size of the decline in L_1 predicted from scenarios A′–E′ (on the order of 30 percent) makes it unlikely that full adjustment could occur in a time span of less than 10 years. Second, a host of governmental and government-sponsored job-support programs, as well as a strict job-tenure law, were instituted and expanded in the 1970s. Assume then that, as of 1977, the economy was adjusted to a 74 percent tax rate (the rate as of about 1973). Depending on the scenario, this should have led to a 15–20 percent decline in L_1 from 1969 to 1977. Since measured unemployment increased only slightly, this leaves roughly a 10 percent discrepancy between measured and predicted employment in the taxed sector as of 1977. Given the aforementioned job-support programs and the

[10] A prediction of future growth rates in Sweden can be derived. Assume that the marginal tax rate in effect today, about 80 percent, remains constant and that the economy is presently adjusted to a tax rate of 75 percent. This assumes a 5-year adjustment lag. Comparing output levels at 75 percent and 80 percent marginal tax rates, something like an 11 percent output decline (±1 percent, depending on the scenario) is to be expected. Spreading this out over the next 5 years, Swedes can await an average growth rate of roughly 1.5 percent over the next 5 years (4.4, less the unexplained 0.75, less 11/5), with yearly growth rates in the neighborhood of 3.5–4.0 percent thereafter.

job-tenure law, the most reasonable use of this 10 percent figure is to interpret it as an estimate of the increase in the level of "hidden unemployment" (including on-the-job leisure) in the taxed sector of the Swedish economy. Since officially measured unemployment is typically only 0.5–2 percent, a 10 percent hidden unemployment figure is not unreasonable.

IV. Concluding Remarks

The analysis here suggests that high taxes on labor income can have quite significant effects on the sale of labor in taxed markets when alternative untaxed uses of labor exist. In this way, taxes can and probably have affected the recent growth performance of the Swedish economy. Further, the treatment permits the conclusion that total tax revenues in Sweden would be maximized at a tax rate of roughly 70 percent, which is 10 percent less than the rate prevailing today.[11]

It is important that the rather severe effects of tax increases above some critical level (70 percent?) may be associated with adjustment lags, at least if the tax increases occur over a short period of time. Social institutions such as the 40-hour workweek and attitudes toward work and labor market participation act as a brake on the adjustment process. Similarly, attitudes and effort devoted toward tax evasion (both legal and illicit) change slowly. Measuring these is perhaps as much the task of the sociologist as the economist. I will refrain from making subjective judgments about whether I think attitudes have changed in the last 10 years in Sweden. There are a number of indications of pressure on the institutions, however: Legislated maternity leaves have been increased; the yearly vacation has been increased from 4 to 5 weeks; there are increasing demands for part-time jobs;[12] and an official commission recently investigated the feasibility of a 30-hour workweek, citing that "the long-run goal [of

[11] If households are assumed to maximize a constant elasticity of substitution (CES) utility function with an elasticity of substitution of 0.5, and the assumptions of scenario A′ are used to derive parameter estimates, the calculated value of t^* is 82 percent. This is close to the present tax rate. Decreasing the elasticity of substitution from 1 (as in the Cobb-Douglas case) to 0.5 takes away about half of the explanatory power of rising taxes as a determinant of growth rates. It also implies a simulated yearly decline in L_1 from 1964 to 1969 of 1.1 percent (the actual rate was 1.6 percent), and implies a 9.6 percent drop in L_1 from 1969 to 1977.

[12] A survey conducted by the university graduates labor union in 1978 and published in the union paper (*SACO-SR tidning*, March 5, 1979) asked, "Do you wish to work part time?" Thirty-one percent answered yes. Since 1975, the percentage of men answering yes rose from 14 percent to 27 percent. A third of all M.D.'s answered yes. Sixty percent of the members surveyed preferred a shorter workweek to increased pay. Forty-four percent of the women wanted a 6-hour day, and 43 percent of the women with full-time jobs wanted to work part time.

the labor unions] is a 30-hour workweek" (Statens Offentliga Utredningar 1976, p. 9).[13]

Appendix

A more detailed description of the parameterization of the model is presented here. Unless stated otherwise, figures apply to 1969.

1. GNP at factor prices was 136.4 billion kronor, whence $Y_1 = 136.4$. Wages were 92.07 (in billions of kronor). In addition, there was an operating surplus of 39.7, some of which is returns to labor. To estimate how much, note that wages paid by industries and households were 62.7. Note also that hours worked by entrepreneurs and employees were 761 and 4,518 million, respectively. The assumption that entrepreneurs have the same wage as employees leads to an estimate of the implicit wages of entrepreneurs of 62.7 × (761/4,518) = 10.56. Thus $L_1 = 92.07 + 10.56 = 102.63$ and $a = 102.63/136.4 = 0.752$.

2. Unfortunately, the only household time-budget study available was for 1976. Activities included in untaxed production are cooking, child care, clothing repair, care of pets, care of dwelling, care of cars, boats, and bikes, care of garden, and other (cleaning, washing, dishes, purchasing, etc.). Time spent in these activities averaged 672 minutes/day. Since the population included in the study consisted of families with children living in apartment buildings, and since the study was conducted during winter months when care of dwelling, garden, etc., are likely to be seasonally low, 20 minutes/day was added to time spent in untaxed activities to seasonally adjust the data. Work (and study) took 642 minutes/day. This implies that $L_2 = (692/642)L_1 = 1.08L_1 = 110.84$. However, taxes were greater in 1976 than in 1969, so the estimate of L_2 for 1969 was that it just equalled L_1, that is, $L_2 = 102.63$.

This estimate of L_2 does not include leisure. To include leisure, note that 647 minutes/day were spent by the average household in the study in recreation and relaxation in 1976. On this basis, L_2 was taken as exactly twice L_1 (i.e., $L_2 = 205.26$) when leisure was included.

The value of untaxed production was estimated by

$$\sum_j n_j \sum_i (h_{ij}^m w_i^m + h_{ij}^f w_i^f),$$

where n_j is the number of households of type j, h_{ij}^m is the number of hours/year spent by males in household j at activity i, w_i^m is the after-tax male wage for activity i, and h_{ij}^f and w_i^f are similarly defined times and wages for females. Six household types were considered: with one adult and 0, 1, or ≥ 2 children, and with two (or more) adults and 0, 1, or ≥ 2 children. The basic data on the n_j were taken from the 1970 *Population and Housing Census*, part 9.

[13] In addition, all of the major political parties in Sweden except the conservatives (with roughly 20 percent of the vote) support a 6-hour working day, the conservatives state only that the matter should be decided in union-management negotiations. One might ask, How does a 30-hour workweek fit with the predictions of the model? A shift from 40 to 30 hours/week is a 29 percent reduction. The actual reduction may be somewhat less if such a change is legislated, however, since labor-market participation might rise (it has risen historically during such changes). Thus something like a 20 percent decline in taxed-sector labor would occur. This is almost exactly the decline needed to restore equilibrium at the 80 percent tax rates prevailing today given the logic of scenarios A′–E′ (cf. Sec. III).

An adjustment was made for the fact that roughly 30 percent of the persons reported to be single adults living with children are actually also living with another adult (for tax and subsidy reasons, it is advantageous to report in this way—the source for the 30 percent figure is the time-budget study, Konsumentverket 1977, p. 30).

Wage data were taken from *Wages, 1969*. The relevant tax rate is the marginal income tax rate plus a correction for income-based subsidies. For 1969, these were 0.478 and 0.052, respectively (for discussion, see Appendix, item 3 in this study). However, application of a 53 percent tax rate is likely to overestimate the taxes which would be paid on marginal income, due to the fact that many of the individuals devoting large amounts of time to home production have low incomes and hence lower marginal taxes than the average income earner. A 40 percent tax rate was thus applied to all untaxed activities except care of pets, care of garden, and other—a 30 percent rate was used for these

The following assumptions were used to estimate hours spent in untaxed production. (1) The two-adult, childless household was considered to have the same uses of time in untaxed production as a two-adult, one-child (ages 4–6) household except that time spent in child care was set at zero. (2) Times in untaxed production for a one-adult, one-child household were 0.94 times the hours for a woman in a two-adult, one-child (ages 4–6) household. The figure 0.94 derives from the consideration that a single parent is likely to spend more time in market work than is an average parent in a two-adult household. (3) The single-adult, no-child household was assumed to have the same uses of time as the average over males and females of time spent in a two-adult, single-child (ages 4–6) household, except that time spent on child care was set equal to zero. (4) The one-adult, ≥ 2 child household was assumed to have the same uses of time as a female in a two-adult, single-child (ages 4–6) household.

In evaluating the value of leisure time, the after-tax "wage" rate was taken as 5 kronor/hour. This figure was generally less than the after-tax wage rates for untaxed productive activities.

3. For each year, the tax-rate base used for calculating tax rates was one plus the payroll tax rate (sources: Industriens utredningsinstitut 1976, pp. 77–78; Svenska arbetsgivareforening 1978). The total tax rate consisted of the payroll tax plus the average or marginal rate on income and the average or marginal indirect tax rate, plus a correction for income-based subsidies, all divided by the rate base. Indirect tax rates were found by dividing total indirect taxes by disposable income and multiplying by 1 minus the average or marginal tax rates on income to find average or marginal indirect tax rates, respectively. It was then assumed that 75 percent of the calculated indirect rates were borne on the margin by labor income. Approximate correction rates for transfers were derived by dividing transfers to households by disposable income. Payroll taxes vary from a few percent in the 1950s to 22.4 percent in 1974 (to 35.1 percent today). Indirect rates, after being multiplied by 0.75, are on the order of 5–10 percent. The transfer correction term rose from 3.5 percent in 1954 to 6.9 percent in 1974.

A sample calculation—for the 1969 aggregate marginal tax rate of 65.5 percent—is as follows. Jakobsson and Normann (1974) report an aggregate marginal income tax rate of 0.478. The payroll tax rate for 1969 was 0.125. Note that this rate is assessed per krona of gross wages (i.e., wages before income taxes) paid by employers. Indirect tax revenues divided by disposable income were 0.209 (= 18,409/87,979), implying a marginal, indirect-tax rate

on net-of-income-tax income of 0.157 (= 0.209[3/4]). In addition, transfer payments divided by disposable income were 0.052 (= 4,553/87,979). Thus, an employer who paid 1.125 kr for labor generated payroll taxes of 0.125 kr and gross income of 1.0 kr. Of the income, 0.478 kr went to income taxes while the remainder (1 − 0.478) was available for consumption and saving and is assumed to have been taxed at a 0.157 rate. The 0.052 transfer rate is treated as an additional marginal tax payment on gross wages. The aggregate tax rate is thus total taxes generated by the 1.125 kr increment in income, divided by 1.125:

$$\frac{0.125 + 0.478 + (1 - 0.478)(0.157) + 0.052}{1.125} = 0.655.$$

References

Adler, H. J., and Hawrylyshyn, O. "Estimates of the Value of Household Work Canada, 1961 and 1971." *Rev. Income and Wealth* 24 (December 1978): 333–55.

Boskin, Michael J. "Efficiency Aspects of the Differential Tax Treatment of Market and Household Economic Activity." *J. Public Econ.* 4 (February 1975): 1–25.

———. "Taxation, Saving, and the Rate of Interest." *J.P.E.* 86, no. 2, pt 2 (April 1978): S3–S27.

Cooter, Robert, and Helpman, Elhanan. "Optimal Income Taxation for Transfer Payments under Different Social Welfare Criteria." *Q.J.E* 88 (November 1974): 656–70.

Douglas, Paul H. "The Cobb-Douglas Production Function Once Again. Its History, Its Testing, and Some New Empirical Values." *J P.E.* 84, no. 5 (October 1976): 903–15.

Feige, Edgar L. "How Big Is the Irregular Economy?" *Challenge* 22 (November 1979): 5–13.

———. "A New Perspective on Macroeconomic Phenomena. The Theory and Measurement of the Unobserved Sector of the United States Economy: Causes, Consequences and Implications." Mimeographed. Univ. Wisconsin—Madison, August 1980.

Gronau, Reuben. "Leisure, Home Production, and Work—the Theory of the Allocation of Time Revisited." *J P.E.* 85, no. 6 (December 1977): 1099–1123.

Harberger, Arnold C. "The Incidence of the Corporation Income Tax." *J.P.E.* 70, no. 3 (June 1962): 215–40.

Hawrylyshyn, O. "The Value of Household Services: A Survey of Empirical Estimates." *Rev. Income and Wealth* 22 (June 1976): 101–31.

Industriens utredningsinstitut. *IUI's långtidsbedomning, bilagor.* Stockholm: Almqvist & Wiksell, 1976.

Jakobsson, U., and Normann, G. *Inkomstbeskattningen i den ekonomiska politiken.* Uppsala: Almqvist & Wiksell, 1974.

Konsumentverket. *Tid och arbete i hushållen.* Stockholm: Allmanna byrån, 1977.

Murphy, M. "The Value of Nonmarket Household Production: Opportunity Cost versus Market Cost Estimates." *Rev. Income and Wealth* 24 (September 1978): 243–55.

Nordhaus, William D., and Tobin, James. "Is Growth Obsolete?" In *Economic Growth*. Economic Research: Retrospect and Prospect, Fiftieth Anniversary Colloquia, vol. 5. New York: Columbia Univ. Press (for Nat. Bur. Econ. Res.), 1972.

SACO-SR tidning, Stockholm (March 5, 1979).

Shoven, John B., and Whalley, John. "A General Equilibrium Calculation of the Effects of Differential Taxation of Income from Capital in the U.S." *J. Public Econ.* 1 (November 1972): 281–321.

Statens Offentliga Utredningar. *Kortare Arbetstid. När? Hur?* Stockholm: Liberforlag, 1976.

Statistiska centralbyrån. *Population and Housing Census, 1970*. Part 9, Stockholm.

———. *Wages, 1969*. Stockholm.

———. *Labor Force Surveys*. Nos. Am 1977:2.4, Am 1978:3.3, Am 1979:32, Am 1979:5.3. Stockholm.

Svenska arbetsgivareförening. *Hur man beräknar företagets sociala kostnader*. Stockholm, 1978.

Comments

A Superior Solution to Captain MacWhirr's Problem: An Illustration of Information Problems and Entitlement Structures

Gene E. Mumy

Ohio State University

> You couldn't tell one man's dollars from another's, he said, and if you asked each man how much money he brought on board he was afraid they would lie, and he would find himself a long way short. [JOSEPH CONRAD, *Typhoon*]

In his novel *Typhoon,* Joseph Conrad posed a classic problem of information revelation. At first glance, the problem appears to be insoluble because of the apparent incentive for individuals to generate wrong information. After further consideration, however, it turns out that an appropriate change in the entitlement structure of the problem allows it to be solved by generating an incentive for each individual to reveal correct information. In this sense, the problem is akin to the class of free-rider problems associated with public goods that can be solved by using the demand-revealing process (Tideman and Tullock 1976).

Conrad's problem is stated in Section I, along with the arbitrary solution imposed by the fictional character Captain MacWhirr. A superior solution, which actually generates correct information, is presented in Section II. Section III contains a brief discussion of the similarities between the superior solution and the demand-revealing process for public goods, followed by conclusions.

The author would like to thank Richard Jensen and William Manson for their helpful comments. They are not, however, responsible for any errors in this paper.

[*Journal of Political Economy*, 1981, vol. 89, no. 5]
© 1981 by The University of Chicago 0022-3808/81/8905-0011$01.50

I. The Problem and Captain MacWhirr's Solution

In *Typhoon*, 200 Chinese workers are being transported to their home province on a ship commanded by Captain MacWhirr, after working for 7 years in various tropical colonies. Each worker's accumulated savings of silver dollars are stored in his own camphorwood chest. As the title of the story suggests, all is not smooth sailing during the trip and the ship runs into a typhoon.

The problem that we are interested in arises when, during the course of the storm, all the wooden chests get smashed and the silver dollars are all scattered between decks. A riot ensues among the Chinese workers as they try to retrieve their money. In order to stop the conflict, the captain sends the first mate and some men to pick up all of the money, which they do. When the storm is over, the captain, being a fair man, wants to return the money to its rightful owners, but he has a problem. He does not know how much money each man had.

The problem here is to obtain information about how much money each man actually had, assuming only that the captain knows the total amount of money and that each man knows the amount he had but not necessarily the amount that anybody else had or even the total amount of money. The captain realizes that he cannot just ask each man how much money he had and then give him that much because some of them might overstate the amount. In this event, the sum of the claims would exceed the total amount of money, thus making a distribution on this basis impossible. The captain concludes that it is impossible to guarantee correct information and therefore imposes an arbitrary distribution that he thinks is fair. According to him, the fair assumption is to assume that all the men had the same amount of money and, therefore, he gives an equal share of the total to each man.

It is obvious that this distribution is not necessarily fair at all unless all the men really did have the same amount of money, which is not known to be the case. The captain's solution to the problem is not a solution to the information problem but rather a surrender. Such a surrender, however, is not necessary, as will now be demonstrated.

II. A Superior Solution

If only the individual knows how much money he had, the captain is right to think that this information can only be obtained by getting the individual to reveal the correct amount. He is wrong, on the other hand, to think that there is necessarily an incentive for the individual to overstate the amount. The mistake here is the implicit assumption that the amount stated by an individual is also the amount to which he

is entitled. Under such an entitlement structure, each individual has a clear monetary incentive to overstate his share. The way to solve the problem, then, is to cut this direct link between amount revealed and entitlement in such a way as to obtain correct information while at the same time entitling individuals to the amounts they actually had.

In Captain MacWhirr's problem, there is a fairly straightforward entitlement structure which generates an incentive for individuals to reveal correct information. Instead of letting the amount claimed be the entitlement, let the maximum amount to which an individual is entitled be the difference between the total amount of money and the amount claimed by everyone else. If an individual claims less than the maximum amount, then he gets what he claims, but if he claims more he is penalized and actually receives less than the maximum amount to which he is entitled. As a result, no individual has an incentive to overstate his share (or to understate it) and he reveals the amount he actually had. This also means that each individual is thereby entitled to the amount he actually had. A more formal presentation of the problem will make this point clearer.

For the following analysis, let r_i = the amount received by the ith individual; c_i = the amount claimed by the ith individual; a_i = the amount the ith individual actually had; and A = the total amount of money. Under an entitlement structure like the one sketched out above, the amount received by the ith individual can be written as

$$r_i = c_i - \max\left\{(1+\alpha)\left[c_i - \left(A - \sum_{j\neq i} c_j\right)\right], 0\right\} \qquad j = 1, \ldots, n, \quad (1)$$

where $A - \Sigma_{j\neq i} c_j$ is the difference between the total amount of money and the amount claimed by everyone else and is, hence, the ith individual's maximum entitlement. It follows that $\alpha > 0$ if individuals are to be penalized for claiming amounts greater than the maximum entitlement.[1]

The individual's problem is to choose c_i so as to maximize r_i. This, however, calls for an assumption about $\Sigma_{j\neq i} c_j$. The reasonable assumption here is that $\Sigma_{j\neq i} c_j = \Sigma_{j\neq i} a_j$, as will be seen by the solution to the individual's problem. If we substitute $\Sigma_{j\neq i} a_j$ for $\Sigma_{j\neq i} c_j$ and note that it follows by definition that $A - \Sigma_{j\neq i} a_j = a_i$, we can rewrite equation (1) as

$$r_i = c_i - \max\,[(1+\alpha)(c_i - a_i), 0]. \qquad (2)$$

[1] If $c_i < A - \Sigma_{j\neq i} c_j$, then $r_i = c_i$ because $(1+\alpha)[c_i - (A - \Sigma_{j\neq i} c_j)] < 0$. Also note that it is assumed that the amount of money that individuals are penalized is kept by the captain.

From equation (2), it is easy to see that $r_i = a_i$ if $c_i = a_i$ but that $r_i < a_i$ if $c_i \gtrless a_i$, and as a result, the individual honestly reveals and receives the amount of money he actually had. Furthermore, this is true for all i, which means that the assumption of $\Sigma_{j \neq i} c_j = \Sigma_{j \neq i} a_j$ is validated and we have a Nash equilibrium solution for our problem.[2]

This solution is obviously superior to the captain's solution because it returns to all individuals the amounts of money they actually had, regardless of what the actual amounts were. This depends on obtaining correct information, which is accomplished by eliminating the direct link between individuals' claims and their entitlements. It is that direct link which gives rise to the incentive for wrong information.

III. Comparison with the Demand-revealing Process and Conclusion

It was stated in the introduction that the superior solution to Captain MacWhirr's problem has a certain similarity with the demand-revealing process for determining preferences for public goods. We are now in a position to see what this similarity consists of. In brief, the similarity is that both are solutions to information problems that are obtained by separating the information revealed by an individual from the individual's entitlement.

The classic public goods problem is that the free-rider phenomenon generates an incentive for individuals to understate their valuations of a public good if their share of the financing depends on that valuation. That is, if individuals feel that their own share of total financing is negligible and, hence, that they are quantity takers from the total collective decision, then they have an incentive to understate their demand for the good if their entitlement (obligation) to share in the financing is determined by the demand they reveal. The twist in the demand-revealing process that causes individuals to accurately reveal their demands is that the entitlement structure is changed from the one that is usually implicitly assumed. Instead of obligating the individual to share only in the financing of the stated quantity desired at a given tax price, the individual is obligated, rather, to share in the financing, at a given tax price, of the quantity chosen by everyone else. This simple change in entitlement basically eliminates the possibility of a free ride. Then, by using suitably designed penalties for deviating

[2] This is not the case with other assumptions about $\Sigma_{j \neq i} c_j$ and, hence, the justification for the assumption used. That is, if it is initially assumed that $\Sigma_{j \neq i} c_j > \Sigma_{j \neq i} a_j$, this means that r_i is maximized with $c_i < a_i$. But, if each person knows that everyone else has the same incentive to act in a similar fashion, this would lead to the conclusion that $\Sigma_{j \neq i} c_j < \Sigma_{j \neq i} a_j$, which contradicts the initial assumption. In the same way, the initial assumption of $\Sigma_{j \neq i} c_j < \Sigma_{j \neq i} a_j$ also leads to a contradiction.

from the basic entitlement given by everyone else, the individual has the incentive to reveal his true demand for the good. Because everyone thus has incentive to reveal his true demand, the total demand for the good is honestly revealed and the appropriate quantity can be chosen. The basic feature of this process is strikingly similar to the basic feature of the superior solution to Captain MacWhirr's problem.

The point here is not that there is necessarily a large class of economics problems that correspond to the fictional problem posed by Conrad and susceptible to solution by the method developed above in Section II. Rather, the fictional problem nicely illustrates that the apparent incompatibility between economic incentives and revealing true information does not mean that the information problem is insoluble. On the contrary, it indicates that a suitable entitlement structure restores the compatibility between incentives and true information, thus allowing the information problem to be solved.

Reference

Tideman, T. Nicolaus, and Tullock, Gordon. "A New and Superior Process for Making Social Choices." *J P.E.* 84, no. 6 (December 1976). 1145–59.

Concentration Changes and Inflation: Some Evidence

E. Woodrow Eckard, Jr.

General Motors Corporation

In a recent article in this Journal, Gisser and Johnson (1979; hereafter cited as G & J) examine the relationship between industrial concentration changes and inflation. They develop an algebraic model relating concentration levels to an economy-wide price index, make assumptions regarding parameter values, and then use the model to simulate a relationship between concentration changes and inflation. The central feature of the model is a Cournot-type specification of the relationship between concentration (or number of firms) and industry price level.[1] Naturally enough, this produces a positive relationship between concentration and the price index and, by extension, between concentration changes and index changes.[2] However, the relationship is weak in the usually observed concentration ranges, and so the authors conclude that "the act of reducing concentration, if successful, is likely to have little impact on the consumer price index" (p. 1377).

The purpose of this note is to present some empirical evidence regarding this issue. In particular, do the facts support the positive, if

Helpful comments from T. G. Marx and T. F. Walton are gratefully acknowledged. Any remaining errors and all opinions expressed herein are the sole responsibility of the author.

[1] Rather than using, say, a four-firm concentration ratio, G & J (1979, p. 1378) assume each industry contains some number of identical firms. Concentration is therefore measured as the inverse of this number.

[2] See Stigler (1968, pp. 36–37) for a description of the Cournot model. However, its theoretical validity, as applied to oligopoly pricing, has been seriously undermined (see, e.g., Stigler 1968, pp. 36–37; Scherer 1970, p. 135). The continued use of this model in industrial organization research appears to be due solely to its mathematical tractability.

[*Journal of Political Economy*, 1981, vol. 89, no. 5]
© 1981 by The University of Chicago 0022-3808/81/8905-0010$01.50

weak, relationship predicted by the G & J simulation? First, a measure of the amount of inflation which can be "associated" with an individual industry is developed. Next, using Census of Manufactures data, this measure is calculated for a sample of four-digit industries over three time periods and then correlated with the corresponding changes in industry concentration.[3] While average economy-wide concentration has been relatively stable in recent years (as noted by G & J), concentration ratios for individual industries have varied sufficiently for identification purposes. The results indicate that concentration changes are unrelated to inflation.

An Inflation Measure

This section develops an empirical inflation measure based on the theoretical model used by G & J.[4] In their model, G & J assume a two-good economy (X and Y) with a single factor of production and no intermediate goods. They could therefore write money income (I) as follows:

$$I = p_x q_x + p_y q_y = VS_x + VS_y, \tag{1}$$

where p is unit price, q is number of units, and VS is value of shipments. However, in a more general model, intermediate shipments occur. Therefore, to avoid "double counting," money income must be written as a function of value added:

$$I = \sum_{i=1}^{N} v_i q_i = \sum_{i=1}^{N} VA_i, \tag{2}$$

where v_i is unit value added (i.e., prime factor cost per unit) and VA_i is total value added for the ith of N goods. The use of unit value added, rather than price, is important in the present context because it eliminates the effect of (nonprime factor) input prices on the industry-specific inflation measure to be developed below. These prices are assumed to be taken as parameters by firms in any particular industry.

A Lespeyres-type "inflation index" can now be defined as follows:

[3] The logic of the market power–inflation relationship implies that the correct empirical specification is between inflation and concentration changes rather than concentration levels. For a discussion of this point, see Stigler (1962, pp. 8–9).

[4] The inflation measure developed here is similar to that of Brainard and Lovell (1966, pp. 860–61). For a discussion of this measure in the context of a simple monetary model of inflation, see Eckard (1978, chap. 2).

$$L = \frac{{}^1I^0}{I^0} = \frac{\sum_{i=1}^{N} v_i^1 q_i^0}{\sum_{i=1}^{N} v_i^0 q_i^0}, \tag{3}$$

where superscripts 1 and 0 refer to current and base years, respectively, and ${}^1I^0$ is a hypothetical nominal income calculated at current-year unit value addeds and base-year quantities.[5] As defined, ${}^1I^0$ is unobservable. However, this can be remedied via some algebraic manipulation. First, rewrite equation (3) as follows:

$$\begin{aligned} L &= \frac{1}{I^0}\left(\sum_{i=1}^{N} v_i^1 q_i^0\right) \\ &= \frac{1}{I^0}\left[\sum_{i=1}^{N}\left(\frac{v_i^1}{v_i^0}\right) v_i^0 q_i^0\right] \\ &= \sum_{i=1}^{N}\left(\frac{v_i^1}{v_i^0}\right)\frac{VA_i^0}{I^0}. \end{aligned}$$

Defining inflation as the change in L during a given period ($\Delta L = L^1 - L^0$), specifying $L^0 \equiv 1$, and recalling equation (2) yields:

$$\begin{aligned} \Delta L &= \sum_{i=1}^{N}\left(\frac{v_i^1}{v_i^0}\right)\frac{VA_i^0}{I^0} - 1 \\ &= \sum_{i=1}^{N}\left(\frac{v_i^1}{v_i^0}\right)\frac{VA_i^0}{I^0} - \frac{\sum_{i=1}^{N} VA_i^0}{I^0} \\ &= \sum_{i=1}^{N}\left(\frac{v_i^1}{v_i^0} - 1\right)\frac{VA_i^0}{I^0}. \end{aligned} \tag{4}$$

This equation states that a given amount of inflation can be partitioned into components associated with each of N elements comprising the economy. Each component is the proportionate increase in that element's unit value added, weighted by its base-year total value added as a fraction of total income.

Abstracting from the effect of relative size (as represented by VA_i^0/I^0), the amount of inflation associated with each element is mea-

[5] Note that this index overstates the "true" price-level change to the extent that (1) the increase in unit value added corresponds to a quality increase and (2) the output vector in period 1 differs from the initial period vector due to consumer responses to relative price changes.

sured by the first term inside the parentheses of equation (4). For convenience, this ratio is defined as follows:

$$R_t = \frac{v_t^1}{v_t^0}. \tag{5}$$

The above is, of course, analogous to a change in average unit price. However, the confounding effects of price changes in inputs purchased from other industries are absent.

Empirical Results

In this section, R_t is calculated for a set of "concentrated" four-digit SIC industries using Census of Manufactures data. This inflation measure is then correlated with concentration changes. But first it is necessary to convert equation (5) to a form which can be calculated from available data. This is accomplished by recalling from equation (2) that $v_t \equiv VA_t/q_t$; therefore:

$$\begin{aligned} R_t &= \frac{VA_t^1/q_t^1}{VA_t^0/q_t^0} \\ &= \frac{VA_t^1/VA_t^0}{q_t^1/q_t^0}. \end{aligned} \tag{6}$$

Value added is available from the Census of Manufactures four-digit industry data.[6] While q_t is not available, a proxy for the ratio q_t^1/q_t^0 can be obtained for certain periods from the Census of Manufactures Indexes of Production. These indexes are estimates of the change in physical output calculated at the four-digit industry level.

Based on these data, R_t is calculated for the time periods 1958–63, 1963–67, and 1967–72.[7] To restrict the sample to "oligopolies," only those industries with four-firm concentration ratios (CR4's) equal to 50 percent or greater at either the beginning or the end of a period are included.[8] The data for the 3 periods are then pooled. To nor-

[6] The census value-added data are at best a rough proxy for "true" value added, i e., the amount of national income originating in each industry. In fact, census value added for all manufacturing industries has consistently overestimated national income originating in manufacturing by roughly one-third (see U.S Department of Commerce 1972, pp. xxxii–xxxiv). Because this error has been fairly consistent over the period covered by the study, we assume that the change in census value added is an unbiased estimate of the change in the "true" value added for each industry

[7] These periods were stipulated by data availability.

[8] The selection of the 50 percent cutoff point is, of course, arbitrary. Its purpose is to avoid a possible bias by excluding those industries which few would argue possess the market power required for the adverse pricing effect postulated under the concentration-inflation hypothesis. The industry concentration ratios used are those published by the Census of Manufactures. Their shortcomings as measures of market

TABLE 1

AVERAGE INDUSTRY INFLATION (R_i) BY CHANGE IN CONCENTRATION FOR THREE POOLED TIME PERIODS (1958–63, 1963–67, and 1967–72)

Changes in CR4	Number of Industries	Average R_i (× 100)	Standard Error
< −10	10	88.5	3.4
−6 to −10	32	91.4	2.7
−1 to − 5	76	93.1	1.3
0	30	92.3	1.7
1 to 5	85	95.9	.9
6 to 10	27	91.2	2.6
> 10	14	89.5	3.6
Total	274	93.1	.7

malize the observations for different "underlying" inflation rates in these periods, each R_i is divided by the increase in the implicit price deflator for the appropriate period.[9] Finally, the data are grouped by the change in CR4.

The results are presented in table 1 (note that the R_i's are multiplied by 100). The data for falling CR4's appear to support the hypothesis that a positive relationship exists between CR4 changes and inflation.[10] Average R_i (× 100) declines from 93.1, for CR4 reductions of 1–5 percent, to 88.5 for CR4 reductions of greater than 10 percent. However, the data for rising CR4's do not support the hypothesis. For CR4 increases of 1–5 percent, average R_i (× 100) is 95.9, while for CR4 increases of 6–10 percent and greater than 10 percent, R_i (× 100) declines to 91.2 and 89.5, respectively. Thus, R_i appears to be more closely related to the absolute value of CR4 change rather than to the direction of change.

This is made clearer in table 2, where the data of table 1 are retabulated by absolute value of CR4 change. For industries with CR4 changes greater than 10 percent in absolute value, average R_i (× 100)

concentration have been well documented (see, e.g., Bock 1972, pp. 11–30). These data are used here because they have come to dominate empirical thinking on concentration, a development best explained by (1) their accessibility and (2) their correspondence to the mass of other data published by the census at the four-digit industry level.

[9] The average industry concentration for the universe of four-digit SIC industries has remained essentially constant between 1958 and 1972 (see, e.g., Allen 1976, p. 665, table 1). Therefore, the inflation rates occurring in each of the three periods must have been due to other (i.e., monetary) phenomena. Removing the effects of these different "underlying" rates eliminates one major source of variance in the pooled data.

[10] The simple correlation between R_i and concentration change for all 274 observations is positive although small in magnitude (r = .024) and statistically insignificant (t = .4).

TABLE 2

AVERAGE INDUSTRY INFLATION (R_i) BY ABSOLUTE VALUE OF CHANGE IN CONCENTRATION FOR THREE POOLED TIME PERIODS (1958–63, 1963–67, and 1967–72)

Absolute Value of Changes in CR4	Number of Industries	Average R_i (× 100)	Standard Error
> 10	24	89.1	2.5
6–10	59	91.3	1.9
1–5	161	94.5	.8
0	30	92.3	1.7
Total	274	93.1	.7

equals 89.1. For industries with absolute concentration changes in the 1–5 percent range, average R_i (× 100) rises to 94.5.[11] While the average R_i (× 100) for the 30 industries with no change in CR4 is less than that for the 1–5 percent change group, it is still greater than the average R_i (× 100) for 6 percent and greater CR4 change groups. Stated alternatively, for the 83 observations with CR4 changes greater in absolute value than 5 percent, average R_i (× 100) equals 90.6. For the 191 observations with CR4 changes of 5 percent or less in absolute value, average R_i (× 100) is 94.2.

What might account for this pattern? One explanation is that competition in concentrated industries (as well as others) often takes the form of cost reductions through productivity increases. Since payroll costs are a substantial portion of value added, productivity increases would be discernible as reductions in unit value added, ceteris paribus. If such cost savings occur among one or more of the top four firms, CR4 rises as the firm (or firms) introducing the cost reduction enjoys a relative increase in sales. If a cost reduction occurs among the smaller firms, CR4 in general falls.[12] However, in either case average industry costs fall. The data in tables 1 and 2 may therefore merely reflect a tendency for firms introducing greater cost reductions to also have greater sales increases, that is, to induce greater CR4 changes.[13]

[11] The simple correlation between R_i and the absolute value of concentration change for all 274 observations is negative ($r = -.18$) and statistically significant at the 1 percent level.

[12] The exception occurs if (1) the sales increase of the smaller firm is sufficient to place it among the top four and (2) the CR4 of the new top four is not less than that of the old. Given the relatively short time periods analyzed (4–5 years), such exceptions are probably infrequent.

[13] Peltzman (1977) has examined the relationship between unit cost changes and concentration changes using a sample drawn from the same universe of census four-digit industries. He found that both concentration increases and decreases are correlated with cost decreases. However, for concentration decreases, the correlation is "statistically insignificant and only a fraction of the effect of similar increases" (p. 245).

One final observation is in order. Since each individual value of R_i is corrected for the appropriate "underlying" inflation rate, an R_i ($\times$ 100) of 100 indicates a "contribution" to the inflation rate equal to that of the economy as a whole, that is, equal to the underlying rate. But, as indicated in table 1, the average R_i ($\times$ 100) for all 274 observations is only 93.1. Thus, the "concentrated" industries in the Census four-digit industry universe, for the given periods, performed significantly better than the remainder of the economy.[14]

Conclusions

In short, the evidence suggests that a policy of *increasing* concentration would be as effective in easing inflation as a policy of decreasing concentration, which is to say, not very effective at all. That is because the concentration-inflation hypothesis which formed the basis for the G & J simulation has the direction of causality between concentration changes and price changes reversed. Concentration changes are induced by corresponding changes in the relative competitive strengths of firms in an industry as manifested by (among other things) price changes, *not vice versa*. In fact, the evidence suggests that neither increasing nor relatively high concentration is associated with "bad" inflationary performance. Policymakers will have to look elsewhere for a cure for inflation.

References

Allen, Bruce T. "Average Concentration in Manufacturing, 1947–1972." *J. Econ. Issues* 10 (September 1976): 664–73.

Bock, Betty. *Dialogue on Concentration, Oligopoly, and Profit: Concepts vs. Data.* Conference Board Report no. 556. New York: Conference Board, 1972.

Brainard, William, and Lovell, Michael C. "Some Simple Propositions Concerning Cost-Push Inflation." *A.E.R.* 56 (September 1966): 857–66.

Eckard, E. Woodrow, Jr. "Industrial Concentration, Market Power, and Inflation: An Empirical Study." Ph.D. dissertation, UCLA Graduate School of Management, 1978.

Lustgarten (1979) has reestimated Peltzman's general model using an alternative specification. He confirmed Peltzman's findings regarding concentration increases but found that "costs and prices rise when concentration falls" (p. 190). However, Peltzman's and Lustgarten's results are not directly comparable to those presented in tables 1 and 2 for two main reasons. First, contrary to the unit value-added cost used above, the Peltzman-Lustgarten cost measure includes material costs and excludes profits. Also, they examine single longer time periods (1947–67 for Peltzman and 1954–72 for Lustgarten) which do not correspond to the three shorter periods of the present study. Nevertheless, the results of both Peltzman and Lustgarten are consistent with the findings of this study in that they do not support a positive relationship between concentration changes and cost-price changes.

[14] This is consistent with previous empirical work on the relationship between concentration levels and inflation (see, e.g., Weston and Lustgarten 1974, pp. 311–13).

Gisser, Micha, and Johnson, Ronald N. "A Note on Inflation and Concentration." *J.P.E.* 87, no. 6 (December 1979): 1377–82.

Lustgarten, Steven H. "Gains and Losses from Concentration: A Comment." *J. Law and Econ.* 22 (April 1979): 183–90.

Peltzman, Sam. "The Gains and Losses from Industrial Concentration." *J. Law and Econ.* 20 (October 1977): 229–63.

Scherer, F. M. *Industrial Market Structure and Economic Performance.* Chicago. Rand McNally, 1970.

Stigler, George J. "Administered Prices and Oligopolistic Inflation." *J. Bus.* 35, no. 1 (January 1962): 1–13.

———. *The Organization of Industry.* Homewood, Ill.: Irwin, 1968.

U.S. Department of Commerce, Bureau of the Census. *1972 Census of Manufactures.* Vol. 1. *Subject and Special Statistics.* Washington: Department of Commerce, 1972.

Weston, J. Fred, and Lustgarten, Steven H. "Concentration and Wage-Price Changes." In *Industrial Concentration: The New Learning,* edited by Harvey J. Goldschmid, H. Michael Mann, and J. Fred Weston. Boston: Little, Brown, 1974.

Book Reviews

The Crisis in Keynesian Economics. By SIR JOHN R. HICKS. Oxford: Oxford University Press, 1974. Pp. 88. £1.75.

My generation accepted the message conveyed by the General Theory with remarkable agility. The analytic interpretation of the message offered by Sir John Hicks effectively contributed to the victorious sweep. But the prospects of a "New Economics" faded within the first 2 decades of the postwar period. The promise of activist stabilization policies failed and eventually yielded a reality dominated by an apparently intractable inflation, protracted unemployment, and volatile exchange rates. Analytic reflections increasingly questioned, moreover, the "Keynesian resolution" of the intellectual dilemma posed by the experiences of the interwar period. These experiences seemed incompatible with inherited price theory. They revealed what appeared to be an immanent flaw of the price mechanism. Two options were available. One would reformulate traditional price theory in order to subsume these experiences within an extended framework, and the other would abandon price theory and replace it with "sociology" or "social psychology." Keynes chose essentially the second option. This was particularly reflected in the treatment of the labor market. It was also reflected in his approach to long-run expectations and their relation to investment expenditures (remember the "animal spirits") This influence pervades the Keynesian establishment and affects much of the "post-Keynesian" literature and the post-Keynesian approach to the inflation problem.

The evolution of monetary theory over the postwar decades may be interpreted to result from a systematic reconsideration of the Keynesian resolution of the intellectual dilemma. It was increasingly argued that an intellectually more productive resolution would continue a viable "classical program." This program is centered on an extended price theory incorporating pervasive information problems but would discard the forms of the "neoclassic synthesis" successfully established during the early postwar phase. This background of our current struggles attaches particular significance to the comments on "the crisis in Keynesian economics" offered by Hicks to the profession.

His material is organized into three chapters: the first covering "the multiplier"; the second attending to money, interest, and liquidity; and the last exploring wages and the labor market. The chapter on the multiplier examines issues badly neglected in the traditional Keynesian literature. Hicks approaches in a somewhat implicit and indirect manner two aspects associated with supply behavior in the context of macroanalysis. The first aspect con-

[*Journal of Political Economy*, 1981, vol 89, no 5]
© 1981 by the University of Chicago

cerns the role of stocks (i.e., inventories) in the multiplier process. The argument elaborates the importance of initial conditions expressed by the inherited pattern of stocks for the magnitude and persistence of the multiplier response to variations in autonomous expenditures. The reader of these passages can hardly avoid wondering about the usefulness of activist fiscal or monetary policy against the background of this chapter.

The other aspect of supply behavior involves the determination of prices. Hicks elaborates his previous differentiation between "fixprice" and "flexprice" market systems. He notes that the traditional multiplier analysis presupposes a fixprice system. The discussion of this problem offers some thoughtful and suggestive remarks to the reader. But it also contains lingering residues of the Keynesian resolution. This appears most particularly in the characterization of the distinct market systems. Flexprices are said to reflect the operation of supply and demand, whereas fixprice patterns move beyond any exposure to market forces. We also read that flexprices are formed by the market and fixprices are set by suppliers (or demanders). This description is quite misleading. All prices are set, and all reflect market forces. The issue centers on the relative frequency of "resetting" and the nature of the market forces shaping the price setting. Several explanations emerged in recent years to account for the relative inflexibilities of prices as a rational market response to specifiable uncertainties and transaction costs. One of the approaches emphasizes that prices are geared in the context of substantial transaction costs to perceived "permanent" market conditions. Price setting disregards under these conditions transitory changes in prevailing conditions. Changes in permanent conditions are moreover perceived only slowly, and prices are thus inflexible relative to the actual but insufficiently perceived underlying market conditions. Such extensions of price theory avoid the Keynesian resolution drifting into ad hoc arguments without a coherent analytic framework.

The second chapter bears on aspects of the Keynesian transmission mechanism and the role of liquidity. The remarks on the latter point are closely related to an analysis which emphasizes that costs of information and transaction differ widely over the spectrum of assets. This variance assures the emergence of "transaction-dominating assets." Future opportunities depend under the circumstances, as Hicks correctly elaborates, on current portfolio compositions. Moreover, the author argues cogently for a wider spectrum of substitution relations surrounding money beyond the narrow choice of "money versus bonds." He still seems to confine, however, the transmission channels to returns on financial assets. The "City Syndrome," so prevalent in the Keynesian traditions, finds it apparently difficult to comprehend that in most countries monetary impulses are dominantly transmitted, in the absence of developed capital markets, via substitutions between money and goods or real assets. Money as the transaction-dominating asset substitutes at the margin in all directions.

The last chapter is particularly informative and suggestive. Hicks discusses the lack of a wage theory in the Keynesian tradition. The reader also infers that Keynes's "wage theorem" opened the gate to a post-Keynesian flood of inflation theories based on more or less explicit denial of economic analysis. Hicks emphasizes in particular the need for an endogenous incorporation of wage push. The formulation couched in terms of "economic" and "noneconomic" causes is singularly misleading, however. On the other hand, the attention given by Hicks to the comparative durability of market relations seems particularly promising for a useful approach to the explanation of

wage behavior. Such reasoning would link the labor market phenomena with evolutions on the output market indicated by the fixprice patterns. The allusions to "fairness" considerations in the wage process seem redundant, however. They have, as they stand, no basis in analysis and add nothing to the substance of Hicks's important suggestions.

The remarks of a thoughtful professional are always worth pondering, and Sir John Hicks certainly deserves our attention. But I wonder about his conclusion that "reformulated Keynes is much more like Keynes than it is like the cruder forms of neo-classical doctrine." I wonder about this statement as intriguing suggestions strewn along the text lead us, as some analytic developments would appear to do, beyond "Keynes" (reformulated or not) and also beyond "neo-classical exercises," crude or sophisticated.

KARL BRUNNER

University of Rochester

The Zero-Sum Society: Distribution and the Possibilities for Economic Change. By LESTER C. THUROW.

New York: Basic Books, 1980. Pp. 230. $12.95 (cloth). New York: Penguin Books, 1980. $4.95 (paper).

This is a book about the economy of the United States, but it could be about almost any other Western developed economy. With perhaps one or two exceptions the advanced countries of the Western world face the same economic problems and have already several years' experience of high unemployment, slow growth, and rapid inflation. They all—with one exception—face special problems arising out of the increased price of oil (the exception, Great Britain, has managed to have its own special problems as a result of *not* having an oil problem and having, instead, a deindustrialization problem caused partly by a high exchange rate). In many cases the problems that Thurow enumerates apply even more acutely to other countries. Some of them are even more dependent on imported oil than is the United States; most of them protect their ailing firms or industries more than does the United States; and most of them spend higher shares of their national product on welfare payments than does the United States. But give or take a problem or two, the picture he paints of the current economic performance of the United States is very familiar in other Western countries, and his diagnosis probably applies in large measure to the other countries as well.

The main theme of the book is that the economic problems facing the American economy today are solvable but only at the cost of somebody losing in the process, and it is almost impossible these days to make the required losses stick on anybody. There are no longer powerless groups in society to whom the burdens of any economic change can ultimately be passed. Civil rights movements, black power, women's lib, poverty lobbies, and so on have become more militant and effective. In addition, there has been increasing Balkanization of the country, and individual states are increasingly reluctant to make sacrifices in the wider national interest—for example, to be the state where the nuclear power stations are to be sited or where the open-cast coal is to be mined.

As Thurow correctly points out, virtually any economic policy decision will affect the distribution of income, and those groups who are likely to be affected are now quick to recognize the threat and to take effective defensive

action. Until recently, changes in the distribution of income between groups that resulted from changes in the economic structure or from government policy measures could be accommodated within a growing economy. Now, however, economic growth has dramatically slowed down, so that we are playing a zero-sum game. Those who lose out relatively to the others also lose out absolutely, and this is more than can be tolerated within the present balance of political power.

Thurow then goes on to show how this general theme applies to some of the major economic problems facing the U.S. economy today. He begins with the energy problem and explains how the free market solution—letting the price of oil rise to world levels—means a large cut in real incomes of certain sections of the population, particularly the poorest. Furthermore, although inflation usually only redistributes income among groups within the economy and does not necessarily impoverish the economy as a whole, it is still greatly resented so that the overall inflationary impact of rising energy prices has to be taken into account as well.

And other measures to alleviate the energy problem—notably more resort to some of the more economically viable energy alternatives such as open-cast coal mining or nuclear power—are opposed by special local interests who are able effectively to delay projects to the point that they kill them off. And, anyway, speculative investment in alternative forms of energy is not attractive to private industry since the high price of imported oil is the result of a cartel, not of corresponding high costs of production. So if some alternative form of energy could be produced that was competitive with oil at, say, $30 per barrel, the oil producers could just cut their price, while still making fat profits, and put the new competitor out of business. Hence, large-scale investment in the search for alternative energy sources must rely largely on the government one way or another. But that means more taxes, and we are all against them. Moral: Whatever way one turns, there seems to be no effective solution in sight to the energy problem without some group or other getting hurt, and all the potential victims are now able to protect themselves.

As regards the problem of inflation, Thurow's analysis follows that of many other commentators and is, in a sense, somewhat inconsistent with much of his basic theme. He spells out the widespread view that inflation is largely the result of certain key groups in society being able to protect themselves against cuts in their real incomes. Since prices and wages are flexible upward but not downward, random changes in various prices and costs will have an upward ratchet effect on the overall price level.

All this is familiar, but, in addition, Thurow introduces into the story an analysis of the labor market which is more original and is the one he has set out in more detail in an earlier book *(Generating Inequality)* and which demonstrates why a cut in demand and an increase in unemployment is likely to have a negligible effect on the rate of inflation. (This is very much in line with many of the econometric macro models on the market these days—not that confirmation from current econometric models is necessarily much comfort to anybody.)

But a major conclusion of Thurow's analysis is that the burden of futile restrictionist monetary or fiscal policies falls mainly on those who are already the weakest economic groups in the economy—such as young people, women, or ethnic minorities. As he puts it: "If a recession is to be used to stop inflation, we draft inflation fighters in a very uneven pattern. In proportion to their size in the labor force, sixteen to twenty-four-year-olds are three times as likely to be drafted as adults. Females are 38 percent more likely to be drafted

than males; blacks are twice as likely to be drafted as whites; and Hispanics are 75 percent more likely to be drafted than whites" (p. 65).

However, in this part of the book Thurow seems to be arguing that society does contain quite a sizable number of "fall guys" who can bear the burden of whatever particular policy society adopts. Hence, what Thurow needs to do in order to ensure greater consistency with his basic theme is to show that the fall guys in the inflation case cannot be made to bear the burden of the adjustment that would be needed in order to solve the other problems. This might not be difficult to do—for example, a rise in gasoline prices would hit white males aged 25–64 as much as, if not more than, automobile owners in other age, sex, or ethnic groups.

Thurow goes on to discuss other current issues, some of which are clear applications of the zero-sum game theme, such as the environmentalist pressures which really involve special pleading by the middle classes to protect those aspects of their standards of life that have moved up in the ranking of their priorities and that can only be fully satisfied at the cost of sacrifices by the poorer groups of the population.

There is also a connection between the zero-sum game theme and Thurow's discussion of the reasons for the slowdown in the rate of growth in productivity in America during the last decade. For he attributes a large part of the slowdown to government policies that prop up ailing firms or industries in one way or another, again because of the effectiveness of political pressures to ensure that nobody suffers from economic change—at least as long as the potential victims are identified in terms of firms, or industries, or cities, or regions, and not in terms of groups of individuals, such as young people or ethnic minorities, for whom the old American virtues of self-reliance and individual initiative must be preserved and protected against the insidious effects of income support schemes.

Throughout the book Thurow repeatedly refers to the overriding importance of society achieving some sort of consensus as to the criteria that should be used to decide which groups should lose out as a result of some policy, and by how much and in what circumstances, and so on. He provides some idea of what sort of criteria of distributive justice might be reasonable, one of the most important of which is that the earnings dispersion of all groups in the economy should be no more unequal than it is now for fully employed white males, who, he shows, have a much smaller dispersion than other groups. Presumably, he would also want to change the relationship between the *mean* earnings of white males and of these other groups, since the latter might not complain about having higher dispersion if their means were twice as high and, conversely, they might not be all that pleased to have zero dispersion if their mean was 10 percent of that of white males.

There are, first of all, some technical objections to Thurow's rather casual use of earnings dispersions for welfare purposes, such as the fact that earners are less than half the population, or that earnings may not be highly correlated with family incomes per capita (a measure that Thurow uses in a different context, although adult equivalent units would have been a better deflator). But the big question is how to achieve the desired dispersion. Thurow advocates some kind of governmental guaranteed job program—not at minimum wages but at the wages appropriate to the skills and so on of the people in question. However, the precise functioning of this program is not spelled out in much detail.

One is likely to be equally skeptical as to the chances of society reaching some sort of consensus on the various distributive objectives that Thurow

enumerates. Distributional criteria have been the subject of debate since Aristotle (if not earlier), who first introduced the concept of "distributive justice" and the notion that people should be rewarded in proportion to their merits. However, as Aristotle pointed out, there is usually much disagreement as to what constitutes a merit. And in the following 2,000 years or more, not much progress has been made in sorting it all out to everybody's satisfaction. Just appointing a couple of philosophers to settle the matter for us will not help; there is a chance that the two appointed would be Rawls and Nozick.

Governments have to obtain some idea—however vaguely—as to what the public might, on the whole, prefer and, in the light also of other considerations of the national interest, act accordingly. If, of course, one is saddled with a government that is incapable of acting since its leaders want to be loved by everybody, it is not surprising that no clear distributive policy is ever adopted.

Of course, Thurow recognizes at one point that, perhaps, he is being naive in supposing that society can solve its problems, and he recognizes that "in the records of history, we certainly would not be the first society that failed to come to grips with its fundamental internal problems" (p. 207). Of course, problems that seem insoluble at any particular point of time—in the sense that the theoretical solutions are not feasible—sometimes become soluble with the passage of time. The actors in the drama may change or learn new lines. Thurow's message might actually get through to some of them. He has an extraordinary gift for clarity and simplicity of exposition, illuminated from time to time with striking phrases, without being technically inaccurate. Such inaccuracies as I have spotted have not been the result of oversimplification but of occasional lapses into carelessness, such as his incorrect attribution to two authors of an article by only one of them (Malcolm Sawyer), or his failure to explain why at one point (p. 11) he says that "welfare" (i.e., income support payments) constitutes only 1.2 percent of GNP in the United States, whereas at another point (pp. 155 and 156) he says that redistributive transfer payments amount to over 10 percent of GNP.

On the whole the diagnosis is probably fairly mainstream though Thurow does well to bring out the zero-sum game element in the situation and their distributional consequences, as well as drawing all the strands together in a more coherent and authoritative exposition than is easily available elsewhere. It is his policy prescription that is the least convincing part of the book. Perhaps there is no solution and never will be. So perhaps we just have to suffer pseudo-controlled experiments, such as the current experiment in testing the British economy to destruction. It remains to be seen whether the U.S constitutional setup can prevent its being subjected to similar experiments.

WILFRED BECKERMAN

Balliol College, Oxford

Autos, Transit, and Cities

A Twentieth Century Fund Report

John R. Meyer and José A. Gómez-Ibáñez

Should Americans be weaned from their automobiles? In this sobering assessment, the authors argue that the automobile is here to stay. $20.00

A Treatise on the Family

Gary S. Becker

One of the world's most creative economists applies economic reasoning to the family.

"This truly path-breaking book . . . is destined to affect the foundations of every science dealing with human behavior."

$20.00 —*Milton Friedman*

Regulation in Perspective

Historical Essays

Thomas K. McCraw, Editor

Leading scholars of business-government relations reassess the American regulatory experience.

Division of Research, Harvard Graduate School of Business Administration

$14.95

Collected Essays on Economic Theory

Volume I: Wealth and Welfare

John Hicks

The first of three volumes by Nobel Laureate Sir John Hicks, this work is a cornerstone of modern consumer theory.

$29.50

The Telecommunications Industry

The Dynamics of Market Structure

Gerald W. Brock

"A stunning piece of work." —*Richard E. Caves*

Brock's study of the evolving structure of the industry focuses on A.T. & T's position and strategy for maximizing its long-run profit and security.

Harvard Economic Studies, 151 $25.00

To order, send check, money order, MasterCard or VISA number, plus $1.50 postage and handling.

Harvard University Press 79 Garden Street Cambridge, MA 02138

An Empirical Model of Labor Supply in a Life-Cycle Setting

Thomas E. MaCurdy

Stanford University and National Bureau of Economic Research

This paper formulates and estimates a structural intertemporal model of labor supply. Using theoretical characterizations derived from an economic model of lifetime behavior, a two-step empirical analysis yields estimates of intertemporal and uncompensated substitution effects which provide the information needed to predict the response of hours of work to life-cycle wage growth and shifts in the lifetime wage path.

Introduction

Over the past several years there has been considerable activity in formulating life-cycle models of labor supply. Most of this work has gone unnoticed in the empirical literature. This study develops an estimable model of labor supply that fully incorporates life-cycle factors, and it devises simple econometric procedures for estimating this model.

Most empirical work on labor supply assumes decision making in a one-period context. Typically, annual hours of work are regressed on the current hourly wage rate and some measure of property income. A worker, however, determines his current labor supply in a life-cycle setting. Unless credit markets are "perfectly imperfect" and there is no human capital accumulation, the supply of labor is a function of

This research was supported by NSF grant no. SOC77-27136 and grant no. 10-P-90748/9-01 from the Social Security Administration. I have benefited from discussions with James Heckman, Robert Michael, and John Pencavel on the material in this paper. I am also grateful for comments from Douglas Brown, William Gould, and Dan Sumner.

[*Journal of Political Economy*, 1981, vol. 89, no. 6]
© 1981 by The University of Chicago. 0022-3808/81/8906-0007$01.50

current and future discounted wage rates as well as wealth and constraints in other periods. Accordingly, regressions of hours of work on current hourly wage rates yield a wage coefficient that confuses the response of labor supply to wage changes of three types: those arising from movements along a given lifetime wage profile, those arising from shifts in the wage profile, and those arising from changes in the profile slope. As a result, the wage coefficient usually reported in empirical studies has no behavioral interpretation in the context of a life-cycle framework.

The theory underlying the model of lifetime consumption and hours of work in this paper represents a natural extension of Friedman's (1957) permanent income theory to a situation in which the relative price of consumption and leisure varies over the life cycle. Theoretical characterizations of consumption and labor supply derived from this theory sharply distinguish between factors determining a consumer's dynamic behavior and factors determining differences in consumption and hours of work across consumers. This separation leads to a manageable empirical model that allows one to discriminate the responses of labor supply to wage changes attributable to movements along a lifetime wage profile from those responses attributable to parametric changes in this profile. In addition, one can estimate the effects of wealth and demographic characteristics on lifetime hours of work.

The organization of this paper is as follows. Section I outlines an economic model of life-cycle behavior. Section II develops and discusses an empirical model of labor supply. Section III interprets the parameters of this empirical model. Section IV contains the empirical analysis.

I. A Life-Cycle Model of Consumption and Labor Supply

The consumer is assumed to choose consumption and hours of leisure at each age to maximize a lifetime preference function that is strongly separable over time, subject to a wealth constraint. Let utility at age t be given by the concave function $U[C(t), L(t)]$, where $C(t)$ is the amount of market goods consumed and $L(t)$ is the number of hours spent in nonmarket activities at age t. The consumer starts life with assets $A(0)$ and operates in an environment of perfect certainty. At each age t he faces a real wage rate equal to $W(t)$ assumed to be exogenously given. The consumer can freely borrow and lend at a real rate of interest equal to $r(t)$ in period t, and his rate of time preference is ρ. A lifetime is assumed to consist of $T + 1$ periods with L^* being the total number of hours in each period.

Formally, the consumer's problem is to choose $C(t)$ and $L(t)$ at each age to maximize the lifetime preference function

$$G\left\{\sum_{t=0}^{T} \frac{1}{(1+\rho)^t} U[C(t), L(t)]\right\} \tag{1}$$

subject to the wealth constraint

$$A(0) + \sum_{t=0}^{T} R(t)N(t)W(t) = \sum_{t=0}^{T} R(t)C(t), \tag{2}$$

where $G(\cdot)$ is a monotonically increasing function, $N(t) \equiv L^* - L(t)$ is hours of work at age t, and $R(t) \equiv 1/\{[1 + r(1)][1 + r(2)] \ldots [1 + r(t)]\}$ is the discount rate which converts real income in period t into its period 0 equivalent with $R(0) = 1$.

Conditions for an optimum are satisfaction of the lifetime budget constraint and

$$U_1[C(t), L(t)] = R(t)(1+\rho)^t\lambda, \qquad t = 0, \ldots, T, \tag{3}$$

$$U_2[C(t), L(t)] \geqslant R(t)(1+\rho)^t\lambda W(t), \qquad t = 0, \ldots, T, \tag{4}$$

where subscripts denote partial derivatives and λ is defined by $\lambda = \lambda^*/G'$, where λ^* is the Lagrange multiplier associated with the wealth constraint (i.e., λ^* is the marginal utility of wealth in period 0) and G' is the derivative of G. According to condition (3), consumption is chosen so that the marginal utility of consumption equals the marginal utility of wealth after adjusting for a discount factor which depends on the rate of time preference and the rate of interest. Condition (4) determines the consumer's choice of leisure. If it is an equality, then a positive amount of labor is supplied to the market. If it is a strict inequality, then all time is devoted to nonmarket activities.

Using the definition of labor supply (i.e., $N[t] \equiv L^* - L[t]$) and the implicit-function theorem, it is possible to solve equations (3) and (4) for consumption and labor supply as functions of the form

$$C(t) = C[R(t)(1+\rho)^t\lambda, W(t)], \qquad t = 0, \ldots, T, \tag{5}$$

$$N(t) = N[R(t)(1+\rho)^t\lambda, W(t)], \qquad t = 0, \ldots, T. \tag{6}$$

The functions $C(\cdot,\cdot)$ and $N(\cdot,\cdot)$ depend only on the functional form of $U(\cdot,\cdot)$. As a consequence of concavity of U and the assumption that consumption and leisure are normal goods, they satisfy

$$C_1 < 0, N_1 \geqslant 0, N_2 \geqslant 0.^1 \tag{7}$$

[1] See Heckman (1974, 1976) for proofs of these inequalities. Heckman develops and uses functions equivalent to those given by eqq. (5) and (6) in his analysis of the behavior of consumption and labor supply over the life cycle.

These consumption and labor-supply functions allow for corner solutions for hours of work either at age t or at any other age t'. No matter what the consumer's labor-force participation pattern over his lifetime, consumption and labor-supply decisions at any age (including the decision to set hours of work equal to zero) are completely determined by the functions $C(\cdot,\cdot)$ and $N(\cdot,\cdot)$ and the values of the variables $R(t)(1+\rho)^t\lambda$ and $W(t)$.[2]

The relationships given by (5) and (6) hereafter will be referred to as the "λ constant" consumption and labor-supply functions. They represent the marginal utility of wealth constant demand functions for consumption and leisure for a particular form of the lifetime preference function given by (1), namely, the one obtained when G is the identity transformation. For this particular choice of G, λ is the marginal utility of wealth in period 0. Given a choice of $U(\cdot,\cdot)$, it is theoretically possible to compute a unique value for λ using data on an individual's consumption, labor supply, and wage rate at a point in time. This fact receives much attention in the formulation of the empirical model which is discussed in the next section.

Substituting the λ constant consumption and labor-supply functions into the budget constraint given by (2) yields the equation

$$A(0) = \sum_{t=0}^{T} R(t)\{C[R(t)(1+\rho)^t\lambda, W(t)]\}. \\ - W(t)\{N[R(t)(1+\rho)^t\lambda, W(t)]\}. \tag{8}$$

This equation implicitly determines the optimal value of λ; λ, then, can be expressed as a function of initial assets, lifetime wages, interest rates, rates of time preference, and "consumer tastes." Concavity of preferences implies

$$\frac{\partial\lambda}{\partial A(0)} < 0 \quad \text{and} \quad \frac{\partial\lambda}{\partial W(t)} \leqslant 0, \qquad t = 0, \ldots, T.^3 \tag{9}$$

[2] If it is optimal for the consumer to work at age t, then condition (4) is an equality and the functions $C(\cdot,\cdot)$ and $N(\cdot,\cdot)$ represent the solutions of eqq. (3) and (4) for the variables $C(t)$ and $N(t) = L^* - L(t)$, respectively. If, on the other hand, the necessary condition given by (4) is an inequality, i.e., $U_2[C(t), L^*] > R(t)(1+\rho)^t\lambda W(t)$, then the consumer chooses not to work. In this case, $N(\cdot,\cdot) = 0$ and the function $C(\cdot,\cdot)$ is the solution of the equation $U_1[C(t), L^*] = R(t)(1+\rho)^t\lambda$ for $C(t)$. In either case, $C(\cdot,\cdot)$ and $N(\cdot,\cdot)$ only contain the variables $R(t)(1+\rho)^t\lambda$ and $W(t)$ as arguments and their functional form depends only on the form of the period t utility function $U(\cdot,\cdot)$. For further discussion on this issue see Heckman and MaCurdy (1980). Introducing age dependence into the utility function does not change any of this analysis. If the utility function at age t is given by $U[C(t), L(t), X(t)]$, where $X(t)$ is a vector of time-varying determinants of "consumer tastes," then a third argument $X(t)$ enters the consumption and labor-supply functions given by (5) and (6). These functions satisfy the restrictions given by (7), and they also allow for corner solutions.

[3] See Heckman (1974, 1976) for proof of these propositions.

Inspection of the λ constant functions reveals that consumption and labor-supply decisions at a point in time are related to variables outside the decision period only through λ. Thus, except for the value of the current wage rate, λ summarizes all information about lifetime wages and property income that a consumer requires to determine his optimal current consumption and labor supply. At any age, any path of wages or property income over a consumer's lifetime that keeps λ and the current wage constant implies the same optimal current consumption and labor-supply behavior.

The λ constant functions represent an extension of Friedman's (1957) permanent income theory to a situation in which the relative price of consumption and leisure varies over the life cycle. According to these functions, current consumption and labor-supply decisions depend on a permanent component and the current wage rate. The variable λ is like permanent income in the theory of the consumption function. At each point in time it is a sufficient statistic for all historic and future information about lifetime wages and property income that is relevant to the current choice of consumption and labor supply. The usual concept of permanent income or wealth does not qualify as a sufficient statistic for this retrospective and prospective information.[4] Given knowledge of permanent income, a consumer also requires information on future wages to determine his optimal current consumption and labor supply. Only if wages are constant over the life cycle, or labor supply is exogenously determined, can λ be written as a simple function of permanent income or wealth.

The λ constant consumption and labor-supply functions fully characterize a consumer's dynamic behavior in a world of perfect certainty.[5] According to these functions, there are two reasons why a consumer might change his consumption or hours of work as he ages: (1) the real wage rate changes, or (2) the rate of interest varies and is not equal to the rate of time preference.[6]

[4] Permanent income here is defined as that stream of income whose discounted value equals the consumer's wealth in present value terms. Formally, permanent income in period 0, Y_p, is defined by the equation $A(0) + \Sigma_{t=0}^{T} R(t)N(t)W(t) = Y_p \Sigma_{t=0}^{T} R(t)$.

[5] See MaCurdy (1978, 1980, in press) for a discussion of the uncertainty case. It is shown in these papers that with minor modifications the empirical specifications of labor supply developed in this section and their implementation in the following sections are consistent with a world in which the consumer is uncertain about his future lifetime path of wages and property income.

[6] Changes in a consumer's tastes can also be a reason for adjustments in consumption and labor supply over the life cycle. As discussed in n. 2 above, if the period utility function is age dependent, the λ constant consumption and labor-supply functions will also be age dependent. It is still true, however, that the λ constant functions fully characterize a consumer's dynamic behavior.

II. An Empirical Model

This section formulates an empirical model of labor supply that is based on the economic model described above. Specific functional forms are proposed for the λ constant labor-supply function and for the relationship between λ and such variables as lifetime wages and initial assets. The following discussion assumes the availability of panel data.

An Empirical Specification for the λ Constant Labor-Supply Function

Assume that consumer i at age t has utility given by

$$U_i[C_i(t), L_i(t)] = Y_{1i}(t)[C_i(t)]^{\omega_1} - Y_{2i}(t)[N_i(t)]^{\omega_2}, \tag{10}$$

where $0 < \omega_1 < 1$ and $\omega_2 > 1$ are time-invariant parameters common across workers, and $Y_{1i}(t)$, $Y_{2i}(t) > 0$ are age-specific modifiers of "tastes." The variables $Y_{1i}(t)$ and $Y_{2i}(t)$ depend on all of consumer i's characteristics which plausibly affect his preferences at age t; these characteristics may include such variables as the number of children present at age t, the consumer's education, and even age itself.

Assuming an interior optimum,[7] the implied λ constant labor-supply function for consumer i at age t in natural logs is

$$\begin{aligned} \ln N_i(t) = \frac{1}{\omega_2 - 1}\{&\ln \lambda_i - \ln Y_{2i}(t) - \ln \omega_2 \\ &+ \ln[R(t)(1+\rho)^t] + \ln W_i(t)\}. \end{aligned} \tag{11}$$

Assuming that "tastes" for work are randomly distributed over the population according to the equation $\ln Y_{2i}(t) = \sigma_i - u_i^*(t)$, the labor-supply function can be written as

$$\ln N_i(t) = F_i + \delta \sum_{k=0}^{t} [\rho - r(k)] + \delta \ln W_i(t) + u_i(t), \tag{12}$$

where $F_i = [1/(\omega_2 - 1)](\ln \lambda_i - \sigma_i - \ln \omega_2)$, $\delta = 1/(\omega_2 - 1)$, $u_i(t) = \delta u_i^*(t)$, $r(0) = \rho$, and we have used the approximations $\ln [1 + r(t)] \approx r(t)$ and $\ln(1 + \rho) \approx \rho$. The unobserved variables σ_i and $u_i(t)$ represent the unmeasured characteristics of consumer i; σ_i is a permanent component, and $u_i(t)$ is a time-varying error term with zero mean. If we assume that the real rate of interest, $r(t)$ for $t \geq 1$, is constant over

[7] Since this study's empirical objective is to examine the labor-supply behavior of prime-age males, this assumption is not unreasonable.

the life cycle and equal to r, then the λ constant hours of work function reduces to

$$\ln N_i(t) = F_i + bt + \delta \ln W_i(t) + u_i(t), \tag{13}$$

where $b = \delta(\rho - r)$.[8]

The intercept term F_i in this equation represents a time-invariant component that is unique to individual i. This study treats F_i as a fixed effect. Since F_i contains $\ln \lambda_i$ as one of its components, one cannot assume that F_i is a "random factor" uncorrelated with exogenous variables of the model. Inspection of equation (8) reveals that λ depends on the values of variables and constraints in all periods. By construction, F_i is correlated with any exogenous variables used to predict a consumer's wages or wealth. Hence, treating F_i as part of the error term would result in biased parameter estimates of the labor-supply function. Treating F_i as a fixed effect, on the other hand, avoids this bias.

Estimating the parameters of equation (12) only requires variables observed within the sample period. Regressions of current hours of work on individual specific intercepts and current wage rates produce a full set of parameter estimates.[9] Because F_i captures the effect of $\ln \lambda_i$, its estimated value summarizes all of the retrospective and prospective information relevant to consumer i's current choices. As there is no need to forecast any life-cycle variables that are outside the

[8] There are other forms of the utility function that have convenient empirical specifications for the λ constant consumption and leisure demand functions. Two such functions are

$$U_i(t) = \kappa_i(t)[C_i(t) + \mu_i^*(t)]^{\omega^*}[L_i(t) + \mu_i(t)]^{\omega},$$

$$\kappa_i(t), \omega^*, \omega > 0, \omega^* + \omega < 1 \text{ or } \kappa_i(t), \omega^*, \omega < 0,$$

$$U_i(t) = \kappa_i^*(t)\,\frac{[C_i(t) + \mu_i^*(t)]^{\omega^*} - 1}{\omega^*} + \kappa_i(t)\,\frac{[L_i(t) + \mu_i(t)]^{\omega} - 1}{\omega},$$

$$\kappa_i^*(t), \kappa_i(t) > 0, \omega^*, \omega < 1,$$

where $\kappa_i^*(t)$, $\kappa_i(t)$, $\mu_i^*(t)$, $\mu_i(t)$, ω^*, and ω are all parameters. Both of these functions are concave. They include Cobb-Douglas, addilog, CES, and Stone-Geary as special cases. The λ constant functions for consumption and leisure are log linear in λ, wages, and the coefficients $\kappa_i^*(t)$ and $\kappa_i(t)$, which represent specific modifiers of tastes. This study uses the utility function given by (10) to formulate an empirical model because it implies a form for the labor-supply function which can be readily compared with labor-supply equations found in existing empirical work.

[9] If utility at age t depends on measured characteristics of the consumer that vary over the sample period, then current values of these "taste-shifter" variables would also be included as regressors. A natural way to introduce such taste-shifter variables is to model the taste coefficient $Y_i(t)$ as a function of the form $\ln Y_i(t) = \sigma_i + X_i(t)\beta - u_i^*(t)$, where $X_i(t)$ is a vector of variables influencing tastes and β is a parameter. For this case, $X_i(t)\beta\delta$ enters as an additional linear term in the λ constant labor-supply equation given by (12) or (13).

sample period, the λ constant functions afford a considerable simplification of the empirical analysis.

Use of the λ constant functions allows one to estimate parameters needed to characterize dynamic behavior without introducing any assumptions regarding a consumer's behavior outside the sample period. To appreciate this point, consider the problem of predicting the additional hours of work a consumer will supply in response to observing a higher wage rate than he observed at a younger age. To obtain an estimate of this response using a traditional model of life-cycle labor supply, one must formally incorporate the worker's future plans in the model. For example, if the worker anticipates an early retirement, future wages corresponding to the retirement years do not influence current labor supply. Thus, the researcher must not include these wages as explanatory variables. Such considerations lead to difficult data requirements and complicated estimation procedures. Using equation (12), on the other hand, it is possible to analyze this problem without knowing anything about a worker's future plans; an individual constant term for each worker accounts for a worker's future plans in a parametrically simple way.

An Empirical Specification for Individual Effects

Estimation of the λ constant labor-supply function given by (12) does not directly estimate all of the parameters required to characterize all aspects of labor supply. Differences among individuals in initial wealth or lifetime wage paths affect the level of hours of work through F_i. To explain any aspect of labor supply other than dynamic behavior (e.g., how the hours of work of two individuals differ at a point in time), one must confront the problem of predicting individual effects.

From the theoretical analysis above, we know that the value of F, or more properly λ, is uniquely determined by the implicit equation given by (8). This equation does not admit an analytical solution for λ given the specific form of the utility function given by (10), even if it is known that this function applies to all ages and the consumer works in each period. The variable λ is a complicated function of initial assets, lifetime wages, the interest rate, the rate of time preference, and parameters representing unobserved "taste" variables.[10] Using such a relationship as an empirical specification is not feasible.

[10] The equation determining λ_i is

$$A_i(0) = \sum_{t=0}^{T} R(t) \left\{ \left[\frac{1}{Y_{1i}(t)\omega_1} R(t)(1+\rho)^t \lambda_i \right]^{1/(\omega_1 - 1)} - W_i(t) \left[\frac{1}{Y_{2i}(t)\omega_2} R(t)(1+\rho)^t \lambda_i W_i(t) \right]^{1/(\omega_2 - 1)} \right\}$$

This study assumes that equation (8) implies a solution for λ in which $\ln \lambda$ can be approximated as a linear function of measured characteristics, the natural log of wages at each age, initial wealth, and an unobserved random variable representing unmeasured characteristics. With this assumption the implied equation for F_i is

$$F_i = Z_i\phi + \sum_{t=0}^{T^*} \gamma(t) \ln W_i(t) + A_i(0)\theta + a_i, \tag{14}$$

where Z_i is a vector of observed variables (e.g., family background variables), a_i is an error term, and ϕ, $\gamma(t)$, and θ are parameters assumed to be constant across consumers.[11] This structural relationship for F_i implicitly assumes that each consumer has a working life of $T^* + 1$ years. According to the theoretical restrictions given by (9), the $\gamma(t)$'s and θ should all be negative.

Unfortunately, to formulate an estimable version of an equation for F_i, we require additional assumptions concerning the forms of the lifetime wage and income paths. In contrast to the λ constant labor-supply function, estimating the parameters of equation (14) requires data which normally are not available. Most variables appearing in this equation are not directly observed, including the dependent variable F_i, wages outside the sample period, and initial wealth. While estimates of F_i are obtained as a by-product from estimating equation (12), we still require a mechanism for predicting wages outside the sample period and initial permanent income. We do this by introducing lifetime profiles for wages and income.

This study assumes that the lifetime wage path is

$$\ln W_i(t) = \pi_{0i} + t\pi_{1i} + t^2\pi_{2i} + V_i(t), \tag{15}$$

[11] The assumption that these parameters are constant across consumers is, of course, only an approximation. Formally, it can be shown that

$$\gamma(t) = \delta\frac{\partial \ln \lambda}{\partial \ln W(t)} = \delta\left[\frac{\partial \ln N(t)}{\partial \ln \lambda} + 1\right]\frac{\partial \ln \lambda}{\partial A(0)} E^*(t),$$

where $E^*(t) = N(t)W^*(t)$ and $W^*(t) = W(t)R(t)$ are the period 0 present value of earnings and wages in period t. This relationship for $\gamma(t)$ is derived by differentiating Roy's identity for $N(t)$ with respect to $A(0)$ to obtain the equation

$$\frac{\partial \lambda}{\partial W^*(t)} = \lambda\frac{\partial N(t)}{\partial A(0)} + N(t)\frac{\partial \lambda}{\partial A(0)} = \lambda N(t)\left\{\frac{\partial \ln N(t)}{\partial \ln \lambda} + 1\right]\frac{\partial \ln \lambda}{\partial A(0)}$$

$$= \frac{\lambda}{W^*(t)}\left\{\frac{\partial \ln N(t)}{\partial \ln \lambda} + 1\right]\frac{\partial \ln \lambda}{\partial A(0)} E^*(t)$$

Since the empirical specification for the λ constant hours of work function implies $[\partial \ln N(t)]/(\partial \ln \lambda) = \delta$, we see that we cannot formally have $\gamma(t) = \delta(\partial \ln \lambda)/[\partial \ln W(t)]$ and $\theta = \delta(\partial \ln \lambda)/[\partial A(0)]$ constant. Notice that the effects of the interest rate and time preference on F_i are absorbed into the coefficients of specification (14).

where π_{0i}, π_{1i}, and π_{2i} are linear functions of the form

$$\pi_{ji} = M_i g_j, \qquad j = 0, 1, 2, \tag{16}$$

M_i is a vector of exogenous determinants of wages which are constant over the consumer's lifetime (e.g., education and background variables), $g_j, j = 0, 1, 2$, are vectors of parameters, and $V_i(t)$ is an error term. This path assumes that wages follow a quadratic equation in age with an intercept and slope coefficients that depend on age-invariant characteristics of the consumer.[12]

Predicting a consumer's initial wealth is complicated by the fact that most data sets do not contain extensive measures of even the consumer's current wealth. Some measure of the consumer's property or nonwage income during the sample period, however, is usually available.[13] Let $Y_i(t)$ and $A_i(t)$ denote the property income and assets of consumer i at age t. If $Y_i(t)$ is the income flow generated by investing assets $A_i(t)$ at a rate of interest equal to r, we have the relationship $Y_i(t) = A_i(t)r$. Assume that the following quadratic equation in age approximates the lifetime path for property income:

$$Y_i(t) = \alpha_{0i} + t\alpha_{1i} + t^2\alpha_{2i} + \nu_i(t), \tag{17}$$

where α_{0i}, α_{1i}, and α_{2i} are linear functions of the form

$$\alpha_{ji} = S_i q_j, \qquad j = 0, 1, 2, \tag{18}$$

S_i is a vector of measured age-invariant characteristics of consumer i (e.g., education and background variables), q_j, $j = 0, 1, 2$, are parameter vectors, and $\nu_i(t)$ is an error term.[14] The intercept α_{0i} can be thought of as a measure of consumer i's permanent income at age 0; that is, $\alpha_{0i} = A_i(0)r$.[15]

Combining the lifetime paths for wages and income with equation (14) creates an equation for F_i that can be estimated using data

[12] In the following analysis it is assumed that this wage equation generates unbiased predictions for lifetime wages. These predictions need not be efficient nor do they need to be the same predictions used by consumers. It is possible to introduce many alternative forms for the wage equation, such as higher-order polynomials or polynomials with other functions of time (e.g., reciprocals) replacing t and t^2, with only minor modifications of empirical specifications used in the following analysis.

[13] These income measures seldom include imputed income generated by consumer durables, which is a major source of property income for most consumers.

[14] In contrast to $W(t)$, $Y(t)$ is determined endogenously in this model. Eq. (17) can be viewed as an approximation to the optimal lifetime path for $Y(t)$ expressed as a function of the exogenous variables of the model.

[15] Formally, this relationship between α_{0i} and $A_i(0)$ is correct only in a continuous time framework. When modeling the problem in discrete time, one must distinguish assets held at the beginning, at the end of the period, and exactly when asset income is earned within the period.

observed within the sample period. Substituting the wage process given by (15) and the relationship $\alpha_{0i} = rA_i(0)$ into equation (14) yields

$$F_i = Z_i\phi + \pi_{0i}\bar{\gamma}_0 + \pi_{1i}\bar{\gamma}_1 + \pi_{2i}\bar{\gamma}_2 + \alpha_{0i}\bar{\theta} + \eta_i, \tag{19}$$

or substituting relations (16) and (18) yields

$$F_i = K_i\psi + a_i, \tag{20}$$

where

$$\bar{\gamma}_j = \sum_{t=0}^{T^*} t^j\gamma(t), \qquad j = 0, 1, 2, \qquad \bar{\theta} = \theta/r, \tag{21}$$

K_i is a vector including all age-invariant characteristics determining either wages, income, or λ (i.e., all the elements of M_i, S_i, and Z_i), ψ is a vector of coefficients, and η_i is a disturbance term which is randomly distributed across workers with zero mean. Equation (19) is a structural relationship between F and the characteristics of a consumer's wage and income profiles. The empirical analysis of this study focuses on estimating the parameters of this equation, ϕ, $\bar{\gamma}_0$, $\bar{\gamma}_1$, $\bar{\gamma}_2$, and $\bar{\theta}$. As we shall see shortly, these parameters have a sound economic interpretation. Equation (20) is essentially a reduced-form equation for (19). By estimating the parameters of this equation, it is possible to predict how F varies across consumers using only age-invariant characteristics of the consumer as explanatory variables.

III. Interpretation of Parameters

In investigating the effect of changes in wages on labor supply, it is important to separate parametric change of the sort usually contemplated in comparative static exercises from evolutionary change due to movement along a life-cycle wage path.[16] A parametric wage change refers to shifts in a life-cycle wage profile (e.g., a shift from path II to path I in fig. 1), while an evolutionary wage change refers to movements along a given profile (i.e., along any path in fig. 1). Thus, parametric wage changes refer to differences in wages across consumers, while evolutionary wage changes refer to differences in wages across time for the same consumer.[17]

Consider the behavior of labor supply over the life cycle. As a consumer ages, he adjusts his hours of work in response to the

[16] This distinction goes back to Ghez and Becker (1975).

[17] This statement is true only in an environment of perfect certainty. If there is uncertainty about the future, a consumer can experience parametric wage changes as he acquires new information about his lifetime wage path. For a discussion of these issues see MaCurdy (1980, in press).

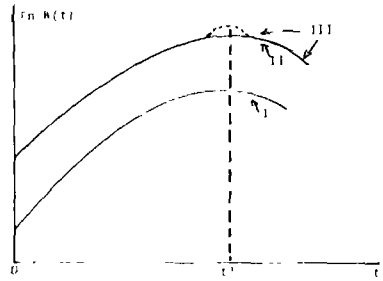


FIG. 1

different wage rates he observes at each point in his lifetime. These labor-supply adjustments represent responses to evolutionary wage changes; they reflect the consumer's desire to supply more hours in those periods with highest wages. There is no wealth effect associated with this kind of wage variation since the wage profile is known to the consumer at the beginning of his lifetime and changes in wages are due only to movement along this given profile. It is apparent from the labor-supply function given by (12) that the value of the parameter δ determines the hours of work response to evolutionary wage changes. Hereafter, I will refer to δ as the intertemporal substitution elasticity. The theoretical prediction for its sign is positive. For the particular form of the utility function given by (10), δ is also the direct elasticity of substitution for hours of work in any two periods.

Now compare the labor-supply profiles of two consumers who face wage paths II and III, respectively. As illustrated in figure 1, the wage profiles for consumers II and III are the same except at age t' when consumer III's wage rate is higher than consumer II's. Let Δ denote the absolute value of this difference in period t' wages. This wage difference represents a parametric wage change because it involves a shift in the lifetime path of wages. It causes the labor-supply profiles for consumers II and III to be different at all ages. Comparing these labor-supply profiles is the sort of problem usually considered in comparative static exercises. In terms of the empirical model outlined above, this higher wage rate has two effects on consumer III's labor supply. The first effect is on the value of F. According to equation (14), consumer III will set a value for F which is lower than the value of F for consumer II by an amount equal to $\gamma(t') \cdot \Delta$. This decline in F implies that at all ages other than t' consumer III's labor supply will be less than consumer II's by a constant fraction. At age t' there is a second effect of the wage difference. Neglecting the decline in F, consumer III's labor supply at age t' will be higher by an amount equal to $\delta \cdot \Delta$. Thus, the total impact on consumer III's hours of work

at age t' is $[\delta + \gamma(t')] \cdot \Delta$. Since $\delta > 0$ and $\gamma(t') < 0$, there is no sign prediction for $\delta + \gamma(t')$, so consumer III's hours of work at age t' may be greater than or less than consumer II's.

The parameters $\gamma(t')$ and $\delta + \gamma(t')$, then, determine the difference in consumer II's and consumer III's labor-supply profiles which is due to the discrepancy in their wage rates at age t'. The quantities $\gamma(t')$ and $\delta + \gamma(t')$ correspond to the usual concepts of cross- and own-uncompensated substitution elasticities. These elasticities describe the response of labor supply to parametric wage changes. They can be used to predict differences in labor supply across consumers. These elasticities do not directly provide information on the response of labor supply to evolutionary wage changes, so they cannot be used to predict differences in a given consumer's labor supply over time. Since the intertemporal substitution elasticity exceeds the own-uncompensated substitution elasticities (i.e., $\delta > \delta + \gamma[t']$), one expects an evolutionary wage change to induce a larger labor-supply response than a comparable parametric wage change. The wealth effect associated with a parametric wage change accounts for the smaller labor-supply response.

Comparing the labor-supply profiles for consumers I and II also involves a parametric wage change. As illustrated in figure 1, consumer II's wage profile exceeds consumer I's by a constant fraction over the entire life cycle. The parametric wage change associated with moving from consumer I to consumer II's wage profile is analogous to increasing the value of the intercept of the lifetime wage path, π_0. This has two effects on labor supply in each period. First, a consumer adjusts his value of F in response to the profile shift. According to equation (19), F declines by an amount equal to $\bar{\gamma}_0 = \Sigma_{t=0}^{T^*} \gamma(t)$ times the increase in the value of π_0. This decline in F implies a fall in hours of work at each age. Second, there is a direct impact on each period's labor supply. Holding the value of F constant, a consumer increases his hours of work by an amount equal to δ times the increase in π_0. The implied total impact on each period's labor supply, therefore, is $\bar{\gamma}_0 + \delta$ times the change in π_0. Because $\bar{\gamma}_0$ is unambiguously negative, there is no sign prediction for this total impact. Since $\bar{\gamma}_0 + \delta$ is less than $\gamma(t') + \delta$, however, the response of labor supply to a shift in π_0 should be less in algebraic value than the response to a shift in the wage profile only at age t'. The wealth effect associated with a shift in π_0 is greater. The labor-supply profile for consumer II, then, can be above or below consumer I's labor-supply profile. It will lie above consumer I's if $\bar{\gamma}_0 + \delta$ is positive.

The empirical specification of life-cycle supply given by equations (12) and (19) provides a convenient framework for estimating the response of labor supply to the different kinds of wage changes

described above. Estimation of the λ constant labor-supply equation produces an estimate of the intertemporal substitution elasticity δ. This elasticity can be used to predict the response of labor supply to evolutionary wage changes; it provides the information one needs to describe a consumer's dynamic behavior. Estimating the equation for F given by (19) produces estimates of the parameters $\bar{\gamma}_0$, $\bar{\gamma}_1$, $\bar{\gamma}_2$, and $\bar{\theta}$. These estimates provide the additional information one requires to predict the response of labor supply to parametric wage and wealth changes, and they can be used to explain labor-supply differences across consumers. Combining the estimates of $\bar{\gamma}_0$, $\bar{\gamma}_1$, and $\bar{\gamma}_2$ with the estimate of δ allows one to predict how labor-supply profiles adjust to changes in the wage-path coefficients π_0, π_1, and π_2. This includes both shifts and slope changes of the wage profiles. The estimate of $\bar{\theta}$ provides the information one needs to predict the response of labor-supply profiles to changes in a consumer's initial permanent income. Estimating the empirical model proposed in this paper, then, fully characterizes a consumer's lifetime labor-supply behavior.

Three Substitution Elasticities

Nowhere in the above interpretation of parameters was there any mention of compensated substitution elasticities. In terms of the above notation, it can be shown using Slutsky's equation that the own- and the cross-compensated elasticities are $\delta + \gamma(t) - E(t)\theta$ and $\gamma(t) - E(t)\theta$, respectively, where $E(t) \equiv N(t)W(t)$ is real earnings at age t.[18] In the analysis of life-cycle behavior, it is important to distinguish sharply compensated elasticities from the intertemporal elasticity (i.e., δ) and the uncompensated elasticities (i.e., own effects $\delta + \gamma[t]$ and cross effects $\gamma[t]$) discussed above.

Some researchers incorrectly infer that intertemporal and compensated substitution elasticities are the same.[19] The fact that an indi-

[18] According to Slutsky's equation,

$$\left.\frac{\partial N(t)}{\partial W(t)}\right|_U = \left.\frac{\partial N(t)}{\partial W(t)}\right|_{A(0)} - N(t)\frac{\partial N(t)}{\partial A(0)}.$$

So

$$\left.\frac{W(t)}{N(t)}\frac{\partial N(t)}{\partial W(t)}\right|_U = \left.\frac{W(t)}{N(t)}\frac{\partial N(t)}{\partial W(t)}\right|_{A(0)} - W(t)N(t)\frac{\partial \ln N(t)}{\partial A(0)} = \delta + \gamma(t) - E(t)\theta.$$

[19] There has been some confusion concerning the interpretation of wage coefficients estimated in labor-supply studies such as those of Ghez and Becker (1975) and Smith (1977), who use synthetic cohort data. Because these studies estimate the response of hours of work to life-cycle wage growth, the wage coefficient is an intertemporal substitution elasticity. While this elasticity constitutes an upper bound for compensated and uncompensated elasticities, it is not one of the familiar elasticities associated with

vidual is at the same level of lifetime utility at all ages in a world of perfect certainty suggests that responses in hours of work to changes in the wage rate over the life cycle represent compensated substitution effects, which in turn suggests that the intertemporal and compensated elasticities are equivalent. These elasticities, however, are used to predict labor-supply responses to different kinds of wage changes. As discussed above, intertemporal elasticities determine hours of work responses to evolutionary wage changes. Compensated elasticities, on the other hand, are like uncompensated elasticities in the sense that they determine responses to parametric wage changes (i.e., wage changes due to shifts in wage profiles rather than movements along these profiles). Compensated elasticities, then, can be used to predict differences in hours of work across consumers whose wage profiles are different and whose lifetime utility is the same.

Formally, the intertemporal substitution effect can also be interpreted as an elasticity that is associated with a particular kind of parametric wage change. In particular, it determines the response of hours of work at age t to a shift in the age t wage rate holding λ or the marginal utility of wealth constant. Thus, whereas uncompensated elasticities hold financial wealth constant and compensated elasticities hold lifetime utility constant, the intertemporal elasticity is equivalent to a parametric wage elasticity that holds λ constant. In the literature on consumer demand, this particular elasticity is known as the specific substitution effect.[20] This correspondence between intertemporal and specific substitution effects is a direct consequence of the assumption that utility is additive over periods.

While the three types of substitution elasticities are related, they are distinct and reduce to the same value only if income or wealth effects are zero. Assuming leisure is a normal good in all periods, these elasticities can be ordered as follows: $\delta > \delta + \gamma(t) - E(t)\theta > \delta + \gamma(t)$; that is, intertemporal responses are greater than compensated responses which are in turn greater than uncompensated responses.[21] It

parametric wage changes. Thus, without imposing nontrivial restrictions on preferences, wage coefficients estimated using synthetic cohort data cannot be used for policy analysis of the sort encountered in predicting labor-supply responses to negative income tax experiments or other proposed changes in tax policies. Elsewhere, Smith (1975) uses these estimated coefficients for exactly this purpose, namely, to predict the effect of income-maintenance programs on hours of work.

[20] See Phlips (1974, pp. 47–50) for further discussion.

[21] We can relate these elasticities using solutions of what is known as the fundamental matrix equation in the literature on consumer demand (see, e.g., ibid.). For leisure demand, it can be shown that: (1) the intertemporal effect is

$$\left.\frac{\partial L(t)}{\partial W(t)}\right|_{\lambda} = \lambda\mu^{22},$$

where μ^{22} is the (2, 2) element of the inverse of the hessian matrix of $U[C(t), L(t)]$, (2)

is also easy to show that intertemporal and compensated responses must be positive, while uncompensated responses may be either positive or negative. Intertemporal and uncompensated substitution effects, then, provide an upper and a lower bound for compensated substitution elasticities.

IV. Empirical Analysis

The λ constant hours of work function given by (12) and the structural equation for F given by (19) provide a manageable empirical model for analyzing labor-supply behavior in a life-cycle setting. This model naturally suggests a two-step estimation procedure. In the first step, one estimates the parameters of the λ constant labor-supply equation. This step provides an estimate of the intertemporal substitution elasticity and all the information a researcher requires to predict how the hours of work of a given consumer will differ at two points in time. In the second step, one uses estimated "individual effects" as dependent variables to estimate the structural equation for F which produces estimates of wealth effects and uncompensated substitution elasticities associated with shifts in the intercept and the slope of the lifetime wage path. This step provides the additional information one requires to predict how labor supply will differ across consumers. Using results from both steps, it is possible to compute the average own- and cross-uncompensated substitution elasticities associated with wage changes in a single year. These results also permit the calculation of upper and lower bounds for compensated elasticities. This two-step estimation procedure exploits the special characteristics of panel data to characterize life-cycle behavior with a minimal amount of computational burden.[22]

---

the compensated effect is

$$\left.\frac{\partial L(t)}{\partial W(t)}\right|_U = \left.\frac{\partial L(t)}{\partial W(t)}\right|_\lambda + \Omega\left[\frac{\partial L(t)}{\partial A(0)}\right]^2,$$

where $\Omega = -\lambda[\partial\lambda/\partial A(0)]^{-1} > 0$; and (3) the uncompensated effect is

$$\left.\frac{\partial L(t)}{\partial W(t)}\right|_A = \left.\frac{\partial L(t)}{\partial W(t)}\right|_U + N(t)\frac{\partial L(t)}{\partial A(0)}.$$

So, if $L(t)$ is a normal good and U is concave, we have

$$\left.\frac{\partial L(t)}{\partial W(t)}\right|_\lambda < \left.\frac{\partial L(t)}{\partial W(t)}\right|_U < \left.\frac{\partial L(t)}{\partial W(t)}\right|_A.$$

The proposed restrictions on labor-supply elasticities follow immediately using this result.

[22] There is an alternative strategy for estimating this structural model of lifetime labor supply. Instead of the procedure suggested above, one could use a one-step procedure. Substituting the right-hand side of eq (19) for F directly into the λ constant

The empirical work reported here uses the randomly designed sample from the Michigan Panel Study of Income Dynamics. The sample consists of observations of 513 prime-age, white, married males for the years 1967–76. Only males continuously married to the same spouse during the period 1968–77 and who were 25–46 years old in 1967 were included in the sample.[23] The labor-supply variable used in the empirical analysis is annual hours of work. The wage variable is average hourly earnings deflated by the Consumer Price Index.

To avoid confusion in this section we must be careful to distinguish between age variables and indexes representing a sample period. As in the previous discussion, t denotes the age of a consumer; in particular, $t = 0$ when the consumer is 25 years old, $t = 1$ when he is 26, etc. The index j, on the other hand, denotes the sample period; $j = 1$ when the observation is from year 1967, $j = 2$ when it is from 1968, etc. The following analysis assumes that there are a total of τ sample periods ($\tau = 10$ for the sample used in this paper). Finally, the notation $t(j)$ denotes the consumer's age in sample period j. Thus, $\ln W_i(t)$ is consumer i's wage at age t, $\ln W_i(j)$ is his wage in period j, and $\ln W_i(t) = \ln W_i[t(j)] = \ln W_i(j)$.

Estimates of the Intertemporal Substitution Elasticities

Estimating the parameters of the λ constant labor-supply equation is simplified by working with a first-differenced version of this equation.

---

labor-supply equation given by either (12) or (13) produces a new equation containing all of the structural parameters of interest. It is possible to estimate this equation in a single step using constrained simultaneous-equation estimation procedures. The two-step procedure offers advantages over this one-step method when there exists misspecification of the equation for F. The presence of any misspecification errors in the F-equation leads to inconsistent estimators for all parameters of the labor-supply equation in the case of the one-step procedure. For the two-step estimation scheme, however, only the second-step estimators for the parameters of the F-equation are inconsistent.

Houthakker and Taylor (1970, chap. 5) and Phlips (1974, pp. 190–93, 250–60) have estimated marginal utility of wealth constant demand functions as an intermediate computational step toward estimating a system of ordinary demand functions. In contrast to their work, I directly estimate the λ constant demand functions as a means of characterizing a consumer's dynamic behavior. I treat λ as a fixed effect which summarizes the effect of historic and future information on current decisions. I thank James Heckman and Orley Ashenfelter for the Houthakker-Taylor reference.

[23] A worker had to satisfy the following criteria as well to be included in the sample: (1) He must be classified as employed or unemployed (i.e., permanently disabled and retired were deleted). (2) Wage and labor-supply data must be available for all years. (3) A worker must report less than 4,680 hours worked per year. The absolute value of the difference in his real average hourly earnings in adjacent years cannot exceed $16 or a change of 200 percent. The absolute value of the difference in the number of hours he works in adjacent years cannot exceed 3,000 hours or a change of 190 percent. The purpose of this last criterion is to minimize difficulties arising from the presence of outliers.

First differencing equation (12) yields

$$D \ln N_i(j) = \delta[\rho - r(j)] + \delta D \ln W_i(j) + \epsilon_i(j), \qquad j = 2, \ldots, \tau, \tag{22}$$

where D is the difference operator, that is, $D \ln N_i(j) = \ln N_i(j) - \ln N_i(j-1)$, and $\epsilon_i(j) = Du_i(j)$ is a disturbance. This equation relates a consumer's labor-supply changes to changes in his wage rate. Notice that differencing eliminates individual effects, and thus it avoids the introduction of incidental parameters. Step one of the estimation procedure is to use this equation to estimate the intertemporal substitution elasticity, δ.

Combining these equations for a given worker into a single system of simultaneous equations creates a model that is particularly well suited for an empirical analysis. Stacking equation (22) for worker i according to the j index yields

$$\begin{bmatrix} D \ln N_i(\tau) \\ D \ln N_i(\tau - 1) \\ \cdot \\ \cdot \\ \cdot \\ D \ln N_i(3) \\ D \ln N_i(2) \end{bmatrix} = \begin{bmatrix} \beta_\tau \\ \beta_{\tau-1} \\ \cdot \\ \cdot \\ \cdot \\ \beta_3 \\ \beta_2 \end{bmatrix} + \begin{bmatrix} D \ln W_i(\tau) \\ D \ln W_i(\tau - 1) \\ \cdot \\ \cdot \\ \cdot \\ D \ln N_i(3) \\ D \ln N_i(2) \end{bmatrix} \delta + \begin{bmatrix} \epsilon_i(\tau) \\ \epsilon_i(\tau - 1) \\ \cdot \\ \cdot \\ \cdot \\ \epsilon_i(3) \\ \epsilon_i(2) \end{bmatrix}, \tag{23}$$

where $\beta_j = \delta[\rho - r(j)], j = 2, \ldots, \tau$. Putting this system into vector notation, we have

$$D \ln N_i = \beta + D \ln W_i \delta + \epsilon_i, \qquad i = 1, \ldots, n, \tag{24}$$

where $D \ln N_i$, β, $D \ln N_i$, and ϵ_i are $(\tau - 1) \times 1$ vectors with $D \ln N_i' = [D \ln N_i(\tau), \ldots, D \ln N_i(2)]$, $\beta' = (\beta_\tau, \ldots, \beta_2)$, $D \ln W_i' = [D \ln W_i(\tau), \ldots, D \ln W_i(2)]$, $\epsilon_i' = [\epsilon_i(\tau), \ldots, \epsilon_i(2)]$, and n is the total number of workers in the sample. The following analysis assumes that the error vectors ϵ_i are independently distributed across individuals once common time effects are removed with the inclusion of year dummies in the labor-supply equations. No restrictions are imposed on the covariance matrix of ϵ_i, which permits arbitrary forms of serial correlation.

The parameters of equation (24) are estimated using standard two-stage and three-stage least-squares procedures which permit the imposition of equality constraints across equations.[24] The wage-growth variables, $D \ln W_i(j), j = 2, \ldots, \tau$, are treated as endogenous variables. The set of instruments used to predict $D \ln W_i(j)$ includes

[24] The two-stage procedures must account for the fact that the covariance matrix of the ϵ_i's is not proportional to the identity matrix when computing standard errors for the estimates.

family background variables,[25] education, age, interactions between education and age, and dummy variables for each year of the sample. Estimation of (24) using simultaneous equation methods takes advantage of the time-series aspect of panel data to estimate the intertemporal substitution elasticity δ with a minimal amount of computational burden. These methods avoid biases arising from pure reporting error in earnings and hours of work, and they offer a flexible framework for testing and estimating alternative functional forms.

Another equation that can be used to estimate δ is one that relates changes in hours of work to changes in earnings. Adding $\delta \cdot D \ln N_i(j)$ to both sides of (24) and solving this new equation for $D \ln N_i(j)$ yields

$$D \ln N_i(j) = \frac{\delta[\rho - r(j)]}{1 + \delta} + \frac{\delta}{1 + \delta} D \ln E_i(j) + \frac{\epsilon_i(j)}{1 + \delta}, \qquad j = 2, \ldots, \tau, \tag{25}$$

where $E_i(j) \equiv N_i(j)W_i(j)$ is real earnings in period j. Stacking these equations creates a model like (24), except that $D \ln W_i$, δ, and ϵ_i are replaced by $D \ln E_i$, $\delta/(1 + \delta)$, and $[1/(1 + \delta)]\epsilon_i$, respectively, and the elements of β become $\beta_j = \{\delta[\rho - r(j)]\}/(1 + \delta), j = 2, \ldots, \tau$. Exactly the same procedures described above for estimating equations (24) are applied to estimate the parameters of the stacked representation of (25). Using the coefficient on earnings, $\delta/(1 + \delta)$, it is possible to construct an estimate of δ.

Table 1 presents estimates of the intertemporal substitution elasticity. Two specifications of the labor-supply equation are considered. One assumes that the interest rate, $r(t)$, is constant over time, and it constrains intercepts in the labor-supply equations (i.e., the elements of β) to be equal over the sample. The other allows $r(t)$ to be different in each period by including dummy variables for each year without any constraints on their coefficients, which permits intercepts to be different each period. These alternative empirical specifications yield similar results. All of the implied estimates of the intertemporal substitution elasticity are positive. According to the estimates of the wage coefficients, δ lies in the range .10–.23. The earnings coefficients indicate a range of .25–.45 for δ. The earnings coefficients indicate a

[25] Family background variables include the education (in years) of both the father and the mother of the consumer and dummy variables indicating parents' economic status at the time the consumer was growing up. Both education and education squared and interactions between these two education variables and age are included as instruments. Coefficients in the "prediction equation" for wage growth are constrained to be equal across time periods. Formal hypothesis tests accept this restriction. In the application of constrained three-stage least squares, the wage-growth equations are not treated as part of the simultaneous-equation system.

TABLE 1

SIMULTANEOUS-EQUATION ESTIMATION OF FIRST-DIFFERENCED LABOR-SUPPLY-EQUATION ESTIMATES OF THE INTERTEMPORAL SUBSTITUTION ELASTICITY

Estimation Procedure	D(Log Wage)	D(Log Earning)*	Intercept	Average of Year Dummies
2SLS	.23	. . .	-.009	. . .
	(2.42)		(4.02)	
3SLS	.14	. .	-.008	.
	(1.97)		(4.26)	
2SLS	. .	.35	-.006	. .
		(2.22)	(4.26)	
3SLS	. .	.25	-.006	. .
		(2.63)	(5.18)	
2SLS	.15	.	. . .	-.008
	(.98)			
3SLS	.10	. . .	. . .	-.008
	(.80)			
2SLS	. .	.45	. .	-.007
		(1.54)		
3SLS	. .	.30	. .	-.007
		(1.67)		

NOTE.—*Absolute values of t-statistics are in parentheses.*

* The estimates and t statistics reported in the "Log Earning" column are for δ; they are computed using the coefficient on earnings, denoted ψ, and its t-statistics. We have $\delta = \hat{\psi}/(1 - \hat{\psi})$. To convert the t-statistics reported for ψ to those for δ requires division by the quantity $(d\delta/\partial\psi) = (1 + \delta)^2$ evaluated at $\delta = \hat{\delta}$.

higher estimate for δ in all cases, but these differences are small relative to their standard errors. The estimates of the intercepts indicate that the real rate of interest exceeds the rate of time preference on average by about 2–4 percentage points.

A comparison of these results with others in the literature is difficult since most studies use cross-section data for their empirical analysis where differences in lifetime wage paths are the primary source of wage variation across observations. As a consequence, they do not estimate the intertemporal substitution elasticity. Estimating equation (13) using cross-section data is complicated by the presence of individual effects, F_i. As discussed above, economic theory implies that F_i is correlated with a consumer's wages and all of his other characteristics. Therefore, it is not reasonable to assume that F_i is a "random factor" uncorrelated with explanatory variables. One, then, cannot directly use observations on individuals from a cross section to estimate the parameters of (13) even if one uses simultaneous equation estimation procedures. Such procedures implicitly treat individual effects as random variables, and this leads to inconsistent parameter estimates. To estimate the intertemporal elasticity using cross-section data, one requires a specification of the labor-supply equation where variation in wages reflects evolutionary wage change.

One approach constructs synthetic cohorts to estimate equation (13) using cross-section data as implemented by Ghez and Becker (1975) and Smith (1977). A synthetic cohort is constructed by computing geometric means of wage rates and hours worked for each age group, and it is assumed to represent the life cycle of a typical individual. The basic assumption underlying this approach is that there are no cohort effects, so that group individual effects (i.e., the average of the F_i's for each age group) are the same for all age groups after adjusting for "smooth" vintage effects. In this case, least-squares estimation of equation (13) using synthetic cohort data produces consistent parameter estimates.[26] This approach allows for measured "taste-shifter" variables such as family size and age. In principle, this approach accounts for the endogeneity of wages and measured characteristics by using group averages as instruments. The problem of treating group effects as random variables does not arise since it is assumed that they are the same for all age groups.

The estimates of the intertemporal substitution elasticity obtained by Ghez and Becker and Smith are comparable to those estimates presented in table 1 of this paper. Becker forms synthetic cohorts using the 1960 U.S. Census, and he obtains estimates for δ ranging from $-.068$ to .44.[27] Smith, on the other hand, treats the family as the relevant decision unit and uses the 1967 Survey of Economic Opportunity to form his synthetic cohorts. He estimates δ to be about .32.[28] Comparing these estimates of δ and those reported in table 1 suggests that cohort effects do not seriously bias estimates based on synthetic cohort data.

Estimates of Responses to Parametric Wage Changes

Estimating the structural equation for F given by (19) provides the additional information we require to predict a consumer's labor-supply response to parametric wage changes. Estimating this equation is not as difficult as it may first appear. It is true that all of the variables appearing in this equation (F_i, π_{0i}, π_{1i}, π_{2i}, and α_{0i}) are not directly observable. But it is possible to construct observable quantities that have expected values equal to these variables. If one replaces the unobserved variables by their observed counterparts, one can employ standard two-stage least-squares procedures to estimate the structural parameters of interest.

[26] Ghez and Becker and Smith do not interpret their parameter estimates as those of a λ constant labor-supply function. Given their log-linear specifications, however, $\ln \lambda$ is absorbed into the intercept of their regression specifications.

[27] Estimates obtained from Ghez and Becker (1975, pp. 112, 114).

[28] Estimates obtained from Smith (1977, p. 244).

First consider the coefficients of the lifetime path for wages. Define the difference operator D_k as $D_k \ln W_i(j) = \ln W_i(j) - \ln W_i(j-k)$. Applying this operator to the wage equation given by (15) and dividing the result by k yields

$$\frac{D_k \ln W_i(j)}{k} = \pi_{1i} + \pi_{2i}[2t(j) - k] + \frac{D_k V_i(j)}{k}.$$

Subtracting $D_1 \ln W_i(2)$ from this equation and dividing the result by $1/(2j - k - 3)$ creates a new equation:

$$\frac{1}{2j-k-3}\left[\frac{D_k \ln W_i(j)}{k} - D_1 \ln W_i(2)\right] = \pi_{2i} + \frac{D_k V_i(j)}{(2j-k-3)k} - \frac{D_1 V_i(2)}{2j-k-3}.$$

Notice that the dependent variable of this last equation has mean equal to π_{2i}. Replacing π_{2i} in the previous equation by this dependent variable allows one to create another observable variable whose mean is π_{1i}. Further substituting these two observable variables for π_{2i} and π_{1i} in the original wage equation allows one to create a third measurable variable whose expected value is π_{0i}. Following this strategy and taking averages to use all the available data, consider the following definitions:

$$\tilde{\pi}_{2i} = \frac{1}{\tau - 2}\sum_{j=1}^{\tau-2}\frac{1}{j}\left[\frac{D_{j+1} \ln W_i(j+2)}{j+1} - D_1 \ln W_i(2)\right] \tag{26}$$

$$\tilde{\pi}_{1i} = \frac{1}{\tau - 1}\sum_{j=1}^{\tau-1}\left\{\frac{D_j \ln W_i(j+1)}{j} - \tilde{\pi}_{2i}[2t(j+1) - j]\right\} \tag{27}$$

$$\tilde{\pi}_{0i} = \frac{1}{\tau}\sum_{j=1}^{\tau}\{\ln W_i(j) - \tilde{\pi}_{1i}t(j) - \tilde{\pi}_{2i}[t(j)]^2\}, \tag{28}$$

where τ is the total number of sample periods. It can be shown that $E(\tilde{\pi}_{hi}) = \pi_{hi}$, $h = 0, 1, 2$. This is an important result because the $\tilde{\pi}_{hi}$'s are observable variables, and it is possible to use the $\tilde{\pi}_{hi}$'s as dependent variables in a simultaneous-equation analysis to estimate the π_{hi}'s consistently.

Similarly, given observations on consumer i's income over the sample period, one can construct the variable $\tilde{\alpha}_{0i}$ using definition (28) with $Y_i(j)$ replacing $\ln W_i(j)$; $\tilde{\alpha}_{0i}$, then, can be used in a simultaneous-equation analysis to predict the intercept of the lifetime income path, α_{0i}, which is a measure of consumer i's initial permanent income.

An analogous strategy can be used to construct a measurable variable to serve as a proxy for F_i. From equation (13) we see that an

average of the quantities $\ln N_i(j) - bt(j) - \delta \ln W_i(j)$ has an expected value equal to F_i, but unfortunately, this average cannot be directly observed since it depends on unknown parameters b and δ. We have estimates of these parameters, however, from the first step of the empirical analysis. A logical alternative to the above average, then, is to form the variable

$$\tilde{F}_i = \frac{1}{\tau} \sum_{j=1}^{\tau} [\ln N_i(j) - \hat{b}t(j) - \hat{\delta} \ln W_i(j)].$$

Asymptotically, $\tilde{F}_i$ has an expectation equal to F_i.

Collecting the above results, we have a complete simultaneous-equations model given by

$$\tilde{\pi}_{hi} = M_i g_h + \eta_h, \qquad h = 0, 1, 2, \tag{29}$$

$$\tilde{\alpha}_{0i} = S_i q_0 + \eta_3, \tag{30}$$

$$\tilde{F}_i = \phi + \tilde{\pi}_{0i}\bar{\gamma}_0 + \tilde{\pi}_{1i}\bar{\gamma}_1 + \tilde{\pi}_{2i}\bar{\gamma}_2 + \tilde{\alpha}_{0i}\bar{\theta} + \eta_4, \tag{31}$$

where the vectors of exogenous variables M_i and S_i and the coefficient vectors g_0, g_1, g_2, and q_0 are defined by (16) and (18), and the η_h's are disturbances. We have one set of equations for each consumer i. The endogenous variables in this model are $\tilde{F}_i$, $\tilde{\alpha}_{0i}$, and the $\tilde{\pi}_{hi}$'s; the exogenous variables are the elements of M_i and S_i; and the structural parameters of interest are $\bar{\gamma}_0$, $\bar{\gamma}_1$, $\bar{\gamma}_2$, and $\bar{\theta}$. The vectors M_i and S_i contain variables determining the coefficients of the lifetime wage and income paths. In the following empirical analysis, they include the consumer's education, his education squared, and family background variables.[29]

To estimate the parameters of the structural equation for F given by (31) consistently, one can employ a standard two-stage least-squares procedure. The standard errors reported by this procedure are valid if the number of time-series observations for each consumer is sufficiently large.[30]

Table 2 presents estimates for the structural parameters of the individual effects equation given by (31), where $\tilde{F}_i$ is computed using three different sets of estimates for b and δ. The estimates of $\bar{\gamma}_0$, $\bar{\gamma}_1$,

[29] Family background variables include the education of both the father and the mother of the consumer and dummy variables indicating parents' economic status at the time the consumer was growing up.

[30] If one does not have a sufficiently large number of these time-series observations, however, the usual standard errors are invalid. The problem lies in the fact that we use estimated values for b and δ to form the dependent variable F_i. In cases where the number of time-series observations is small, one must adjust the usual standard errors to account for errors in estimating b and δ. The precise form of this adjustment can be obtained from the author upon request. While this adjustment is not complicated, it does require the use of matrix operations. This adjustment was very minor in every instance for which it was used in this study, which suggests that even 10 observations per person is large enough to neglect making any adjustments in standard errors

TABLE 2

SIMULTANEOUS-EQUATION ESTIMATION OF FIXED-EFFECTS EQUATIONS

$\hat{b}$,	$\hat{\delta}$	$\bar{\gamma}_0$	$\bar{\gamma}_1$	$\bar{\gamma}_2$	$\bar{\theta}$	Intercept
−.009,	.1	−.05	−.83	−10.47	−.00026	7.81
		(1.5)	(1.4)	(1.2)	(.48)	(157)
−.009,	.15	−.07	−1.08	−13.03	−.0001	7.77
		(1.78)	(1.65)	(1.31)	(.16)	(137)
−.009,	.23	−.10	−1.46	−16.86	.0001	7.68
		(2.13)	(1.9)	(1.4)	(.2)	(119)

NOTE.—Absolute values of t-statistics are in parentheses. The dependent variable $\tilde{F}_i$ (i.e., the proxy variable for F_i) is computed using the estimates $\hat{b}$ and $\hat{\delta}$ reported in the stub column. Income is measured in thousands of dollars.

and $\bar{\gamma}_2$ are all negative as theory predicts. These estimates decrease monotonically as one uses a higher estimated value of δ to compute the proxy variable for individual effects, $\tilde{F}_i$. Since uncompensated substitution elasticities associated with permanent wage changes are calculated by adding the estimate of δ to the estimate of $\bar{\gamma}_0$, an inverse relationship between the estimates of δ and $\bar{\gamma}_0$ is required if uncompensated permanent wage elasticities are to remain constant for different choices of $\hat{\delta}$ in computing $\tilde{F}_i$.

All of the estimates of the initial permanent income coefficient, $\bar{\theta}$, are statistically insignificant and very small. The measure of property income (i.e., $Y_i[t]$) used in this empirical analysis is total family income minus husband's earnings in thousands of 1967 dollars. While the estimates of $\bar{\theta}$ are negative as theory predicts when low estimated values of δ are used to construct $\tilde{F}_i$, we see that the estimated effects of a change in income on hours of work are minute; a $1,000 increase in initial permanent income leads to at most a .026 percent decrease in hours of work. Several other measures of property income were tried in the empirical analysis;[31] in every case the estimates obtained for $\bar{\theta}$ were of the order of magnitude reported in table 2.

Combining estimates of $\bar{\gamma}_0$ and δ allows one to form estimates of cross- and own-uncompensated substitution elasticities and bounds for compensated elasticities associated with wage changes in a single year. Dividing the estimate of $\bar{\gamma}_0$ by the length of the working life produces an estimate of the average cross-uncompensated elasticity. Using results from the second row of table 2 and assuming a working life of 40 years implies a cross elasticity equal to −.0018. Adding the estimate for δ (which is .15 for the second row of table 2) to this cross elasticity implies a value of .15 for the average own-uncompensated

[31] One measure of income excluded wife's earnings. Others included imputed income for house ownership. Due to data limitations, it was not possible to include imputed income from other forms of consumer durables, which certainly constitute a major component of a consumer's nonwage income.

elasticity. Since the average own-compensated elasticity lies between the intertemporal and the average own-uncompensated elasticity, we conclude that this compensated elasticity is also approximately equal to .15. Increasing a consumer's wage rate in period t by 10 percent, then, leads to about a 1.5 percent increase in his hours of work in period t and approximately no change in his hours of work at other ages.

Combining estimates of $\bar{\gamma}_0$, $\bar{\gamma}_1$, and $\bar{\gamma}_2$ with estimates of δ provides the information needed to predict a consumer's labor-supply response to shifts in his wage profile. In response to a uniform 10 percent increase in wages at all ages (i.e., a parallel shift in the log wage profile), the estimates of the second row of table 2 predict that a consumer will adjust his hours of work by an amount equal to $(\hat{\bar{\gamma}}_0 + \hat{\delta})10\% = (-.07 + .15)10\% = .8$ percent at all ages. There is, then, a small positive response in a consumer's labor supply to parallel shifts in his log wage profile. If the slope of a consumer's wage profile is altered by changing the coefficient on the linear term (i.e., π_{1t}) by Δ percent, the estimates for $\bar{\gamma}_1$ and δ indicate that his hours of work at age t change by $(\hat{\bar{\gamma}}_1 + \hat{\delta}t)\Delta\% = (-1.08 + .15t)\Delta\%$. Recall that t here measures a consumer's age and takes a value of 0 when the consumer is 25 years old. Hence, hours of work decline at early ages (i.e., prior to age 32) and they increase at later ages in response to this sort of increase in the slope of the wage profile. The same is true when the slope of the wage profile is altered by changing the coefficient on the quadratic term (i.e., π_{2t}) by Δ percent. Hours of work at age t adjust by an amount equal to $(-13.03 + .15t^2)\Delta\%$, which also implies a decline in hours of work at younger ages (i.e., prior to age 34) and an increase at older ages.

The estimates above of uncompensated elasticities are generally consistent with results found in cross-section studies and the popular notion that the lifetime labor-supply curve of prime-age males is not very responsive to permanent wage changes. If, in a cross-section analysis, one purges wages and income of their transitory components using a simultaneous-equation method and interprets the estimated coefficients as some sort of "lifetime average" relationship, then one finds small positive estimates for wage coefficients and negative or zero estimates for income coefficients for prime-age males,[32] which agrees with the results above. In contrast to previous work, however, the estimates above also indicate that to predict the response of labor

[32] Accounting for the endogeneity of $\ln W_i(t)$ and $Y_i(t)$ has been shown to have a significant effect on cross-section estimates. DaVanzo, Detray, and Greenberg (1976) find that treating both wages and income as endogenous variables leads to positive estimates for wage coefficients and negative or zero estimates for income coefficients for prime-age men, which are consistent with the empirical results reported above. Neglecting this endogeneity produces estimated coefficients with opposite signs.

concentrate on the intermediate case where the intertemporal pattern of expenditures is unchanged. In this case, at a given value of r, the decline in current C^d net of the increase in Y^s exactly matches the rise in $\bar{G}$.[3]

The Role of Public Services

The government is viewed as utilizing its commodity purchases at each date, G_t, to provide a contemporaneous flow of public services to the private sector. These services are treated as provided free of charge to household/producers. Two types of services are considered. One form is modeled as a direct conveyer of utility to households. Examples which do not encompass the traditional roles of government include parks, libraries, school lunch programs, (subsidized) hospitals, and, possibly, highway and transportation programs. (The latter category could be viewed alternatively as an input to private production functions.) An important feature of these forms of public services is the possibility of close substitution with private consumer spending.

The second type of service is an input to private production processes, which can apply either to businesses or households. Examples include the provision of a legal system, aspects of national defense, fire and police services, education, and various regulatory activities. (The last item is likely to exhibit negative marginal product.) In some cases these services would be close substitutes for private inputs of labor and capital. However, in areas like the provision of a legal system and national defense, the public services are likely to enhance the marginal products of private factors.

In many situations a particular government activity would exhibit features of both general types of public services that are being considered. The extent to which each feature was represented would vary across a wide range of programs. Despite this real world diversity, the formal analysis proceeds as if there were a single type of governmental activity, which has service attributes that are partly of the direct-utility type and partly of the productive-input type.

Government-provided services are often modeled as "public," as opposed to private, goods in the sense of being "nonrival"—one person's enjoyment of the good does not diminish the enjoyment by another person.[4] It is doubtful that this characteristic applies to the

[3] This result does not depend on an infinite horizon for the representative household. Finite lives can alter the effects of some government actions, such as changes in public debt or social security, that involve a publicly mandated shift in incomes across generations. Government purchases that are financed contemporaneously by taxes (which are independent of age) do not involve these considerations.

[4] This characteristic is embedded in the theoretical analysis of Samuelson (1954).

majority of government purchases at the current time. Falling outside of this category would be the bulk of expenditures on education, hospitals, school lunch programs, and any service that is subject to congestion, such as parks, courts, libraries, transportation and highway projects, and police/fire services. Even in the case of national defense, the benefits to individuals are likely to be relative to the total amount of property that is being defended, because the level of external threat would respond to the potential prize from conquest (see Thompson 1974).

In a nonrivalry situation, individual utility or production would depend on the total of government services rather than on the quantity provided to the particular economic unit. Because nonrivalry seems atypical, the modeling assumes that individual utility and production depend on real government purchases per capita. However, the general form of the analysis would not be altered appreciably if some elements of nonrivalry were introduced.

Consider, first, the direct interplay in utility functions between government services and household choices of consumption and leisure. Suppose, as stressed by Bailey (1971, chap. 9), that the contemporaneous levels of (per capita) G_t and (individual) C_t are close substitutes in utility terms. For example, assume that each unit of G_t (per capita) is viewed as providing utility services that are equivalent to a fraction θ of a unit of contemporaneous individual consumption expenditure.[5] That is, household utility depends on the effective consumption flow at each date, $C_t^* \equiv C_t + \theta G_t$, where $0 \leq \theta \leq 1$. The formulation neglects this type of utility substitution among noncontemporaneous values for C and G or between G and leisure.[6] The provision of these types of public services means that households obtain units of effective consumption, C^*, that exceed the quantity of private real expenditures, C. The permanent flow of government purchases, $\bar{G}$, can be used to finance the uniform effective consumption flow, $\theta\bar{G}$.[7] This aspect of public services offsets the negative permanent income effect from G that was described earlier (see also n. 6 above). The permanent income effect that is pertinent to private choices on consumption expenditure and leisure is now $(\theta - 1)\bar{G}$—the condition $0 \leq \theta \leq 1$ implies that the permanent income effect of $\bar{G}$ is still less than or equal to zero, but no larger than one in magnitude.

[5] The parameter θ can be viewed in the following analysis as applying to the marginal unit of G_t.

[6] The time path of G_t could also affect overall household utility in a form that was additively separable from the time paths of effective consumption and leisure. In this sense the θ parameter need not limit the utility value that households attach to public services. This possibility does not invalidate the subsequent discussion of permanent income changes that are induced by shifts in $\bar{G}$.

[7] The discussion assumes that the inequality constraint, $C_t \geq 0$, is never binding.

With $\bar{G}$ held fixed, an increase in G_t now implies some direct crowding out of contemporaneous private commodity demand. In order to maintain the level of effective consumption, C_t^*—which is appropriate when $\bar{G}$, the other determinants of permanent income, and r are held fixed—C_t^d must decline with G_t in accordance with the parameter θ. The greater the utility substitution (at the margin) between C_t and G_t—as measured here by θ—the larger the negative response of C_t^d to an increase in G_t. As long as $\theta \leq 1$ applies, aggregate commodity demand for date t—$Y_t^d \equiv C_t^d + G_t$—rises as the nonnegative fraction, $(1 - \theta)$, of increases in G_t (when $\bar{G}$ is held fixed).

Consider next the role of public services as an input to private production processes. It is assumed that public services of this type have a positive marginal product, which is denoted by MPG. The condition MPG ≤ 1, which is assumed to hold, implies that the marginal response of (aggregate) private output to an increase in (aggregate) G does not exceed the social cost of providing the extra public services input. Through its role as a productive input, an increment in G raises commodity supply, Y^s, for given levels of private factor inputs. Note that part of total output will be utilized to provide intermediate goods, which take the form of publicly supplied production inputs. Although there are good reasons in principle for deleting these intermediate goods from measures of final product, this approach is not followed in the national accounts.[8] In the theoretical analysis the output measure, Y, is also gross of this type of intermediate production.

Because the public-service inputs are provided freely, a change in G alters private sector real incomes in accordance with the marginal product, MPG. The representative household receives a per capita share of this extra real income. This effect further offsets the inverse influence of $\bar{G}$ on permanent income. The net effect now depends on the term $(\theta + \text{MPG} - 1)$, which is nonpositive but no greater than one in magnitude if $0 \leq \theta + \text{MPG} \leq 1$ applies.[9]

If variations in G alter the marginal products of private productive inputs, then additional effects would arise through changes in factor demand functions. Cases where public services substitute for private inputs—such as the provision of public rather than private

[8] These matters are discussed in Kuznets (1948, pp. 156–57) and Musgrave (1959, pp. 186–88). The double-counting property for publicly provided production inputs implies that empirical counterparts of total output, like real GNP, overstate the response of final output to government purchases.

[9] Recall that the analysis deals with a composite government service that has attributes of the direct-utility (θ) and production-input (MPG) type. A particular category of purchases is unlikely to exhibit a high value for both parameters, θ and MPG. Therefore, if $\theta \leq 1$ and MPG ≤ 1 apply for each category of purchases, the condition $\theta +$ MPG ≤ 1 is likely to hold for the composite over all categories.

guards—would generate reductions in the marginal product of labor. On this count, the private demand for labor would tend to fall when G rises. In some other cases the private demand for capital services would decline. However, to the extent that background services like national defense and a legal system are expanded, factor marginal products are likely to rise, which would generate the opposite responses in private factor demands. The main analysis neglects the array of possible effects of government purchases on private factor marginal products.

Commodity-Market Clearing

The equilibrium condition for the commodity market is given by

$$Y^d \equiv C^d(\ldots, \underset{(-)}{r}, \underset{(-)}{G}, \underset{(-)}{\bar{G}}) + G = Y^s(\ldots, \underset{(+)}{r}, \underset{(+)}{G}, \underset{(+)}{\bar{G}}). \tag{3}$$

Recall that the analysis has neglected the investment component of demand and has, thus far, assumed lump-sum taxation. Time subscripts have been omitted for convenience. Signs beneath the independent variables refer to partial derivatives. The omitted arguments in the C^d and Y^s functions involve various fixed aspects of households' permanent incomes, the production technology, and so on. As mentioned before, the real rate of return, r, exerts intertemporal substitution effects that are negative on C^d and positive on Y^s. A rise in G has a negative crowding-out effect on the contemporaneous choice of C^d. The impact of G on Y^s is nonnegative if MPG ≥ 0. A rise in $\bar{G}$ reduces the pertinent measure of households' permanent incomes, as discussed before, which leads to a decrease in C^d and a rise in Y^s.

Effects of a Temporary Rise in Government Purchases

Consider a temporary expansion of real government purchases, where G rises while $\bar{G}$ is held fixed. The budget condition from equation (1) implies that these purchases are financed by a contemporaneous increase in real taxes-net-of-transfers, T. In fact, for the context of a temporary rise in government spending—which is most pronounced at the federal level during wartime—it is more natural that the bulk of contemporaneous finance would take the form of interest-bearing debt issue rather than tax increases. This behavior allows the government to spread the higher taxes necessitated by temporary spending over a large time interval instead of implementing exceptionally high tax collections for a few periods (see Barro [1979, 1980*b*] for discussions). With $\bar{G}$ held fixed, it would be possible to utilize interest-bearing debt so as to maintain the entire initial time

path of net real tax collections, T_t. In any event, for the setting of lump-sum taxes, the present value of tax obligations is not altered by shifts between public debt and taxes (see n. 2 above). Since households' calculations of permanent incomes depend only on this present-value magnitude, the effects of increases in government purchases that are financed by debt issue would coincide with the effects of those that are financed by higher taxes.

The rise in G reduces C^d on the left side of equation (3) in accordance with the utility substitution parameter, θ. Therefore, aggregate demand, Y^d, rises on net by $(1 - \theta)$ times the increase in G. If the marginal product of public services, MPG, is positive, Y^s rises with G on the right side of equation (3). Overall, the shift in "excess demand," $Y^d - Y^s$, is determined by the term $(1 - \theta - \text{MPG})$. It was already assumed that this term is nonnegative—that is, the direct, one-to-one effect of G on aggregate demand is offset only partially by the utility-substitution and productive-input aspects of government purchases.

Since an increase in G raises excess commodity demand, a rise in the real rate of return is required in order to restore commodity-market clearing.[10] This response in r reduces C^d and raises Y^s. The rise in Y^s reflects the substitution of current work effort for planned future effort. Since Y^s was also increased directly by the rise in G, it is apparent that equilibrium output rises. This output effect is greater the smaller is the value of θ, the larger is the value of MPG, and the greater is the real interest rate elasticity of Y^s relative to that of C^d. In the polar case where $\theta = 1$ and MPG $= 0$, the response of Y and r to G would be nil. In this circumstance, government purchases would amount to lump-sum transfers to households, because G_t and C_t were perfect substitutes in the utility function.

Private consumer spending, C^d, is crowded out from the rise in r and from the initial negative effect of G. Therefore, the positive response of Y to G must be less than one-to-one—that is, the model exhibits an output dampener rather than a multiplier.

The positive response of output to temporary movements in government purchases would apply especially to wartime periods.[11] The higher real rate of return can be viewed as a price signal that induces the intertemporal substitution of resources toward periods such as wars in which aggregate output is valued unusually highly. This type

[10] Because the expansion of G is temporary, the rise in the equilibrium real rate of interest would also be temporary. An extension to allow divergences between current and expected future real rates of return does not alter the basic analysis.

[11] Wartime may also be associated with uncertainties on maintaining property rights, which would tend to reduce private investment demand. The possibly changing probability of winning or losing a conflict would enter in this context. The analysis abstracts from these effects and from controls on prices or interest rates. Also excluded are effects of patriotism or coercive behavior, such as conscription

of substitution has been stressed by Hall (1980, Sec. 2), who points out also that this behavior differs in some important respects from the responses of supply to monetary misperceptions that occur in some business cycle theories that stress intertemporal substitution on the supply side (e.g., Lucas and Rapping 1969; Lucas 1975; Barro 1980*a*). The effect of temporary government purchases on the time arrangement of work and production does not rely on elements of misperception with respect to the general price level or other variables.

Effects of a Permanent Rise in Government Purchases

Suppose now that G and $\bar{G}$ rise by equal amounts. The increase in G was shown already to raise excess commodity demand in accordance with the term $(1 - \theta - \text{MPG})$. A higher value of $\bar{G}$ was shown earlier to reduce effective permanent income by the same factor. It was also noted above that the typical or intermediate response of C^d net of Y^s would be one-to-one with this type of change in effective permanent income. (This response arises when households reduce expenditures on consumption and leisure by the same amount in each period.) In this case the response of $(C^d - Y^s)$ equals $-(1 - \theta - \text{MPG})$ times the change in $\bar{G}$. Since this response exactly offsets the excess demand effect of G, the overall response in $(Y^d - Y^s)$ would be zero. It follows in this case that the real rate of interest is invariant with permanent changes in government purchases (under lump-sum taxation).[12] Although the invariance of r would not hold in general, this conclusion for an intermediate case contrasts with the presumption of a positive effect for the setting of a temporary increase in government purchases.

When r is unchanged, it follows from equation (3) that consumption falls, and total output rises. The expansion of production reflects partly the direct effect of G on Y^s (which was assumed to be positive) and the negative income effect of $\bar{G}$ on leisure. (The net response of leisure becomes ambiguous when income taxation is introduced—see the subsequent analysis.) The decline in consumption means that output rises by less than one-to-one with the permanent expansion of government purchases; that is, an output dampener is again predicted.

Non-Lump-Sum Taxation

Some of the results are affected by the unrealistic assumption that government expenditures are financed by lump-sum taxation. This

[12] It follows also—assuming no direct effects on the marginal product of capital schedule—that capital accumulation would be unaltered

section explores the consequences of income taxes in a simple environment. Because the present analysis is concerned only with tax effects that are systematically related to changes in government purchases—rather than with public finance questions, per se—this simplified analysis may be adequate.

Suppose that government expenditures are financed by a general income tax. Let τ_t represent the effective tax rate on incomes that accrue during period t.[13] In a nonproportional tax setup, τ_t would represent the average marginal income tax rate. Assume now that the government can borrow and lend at the real rate of interest, r—the same rate that applies to the private sector. Given the permanent flow of real government spending, which includes $\bar{G}$ and the comparable measure for real transfers,[14] and the possibility for variations in public debt, there exists some income tax rate $\bar{\tau}$ that is constant over time and also just satisfies the government's intertemporal budget constraint.[15] Although many patterns of time-varying tax rates would also satisfy the government's budgetary requirements, it will be desirable in most circumstances[16]—in terms of minimizing the distortions that are imposed on the private economy—to stabilize income tax rates over time. That is, the government would adjust its public debt issues and redemptions in order to prevent divergences between current and expected future income tax rates. (For discussions of this type of result, see Barro [1979, 1980*c*]; Kydland and Prescott [1980, pp. 185–86].) In this setting changes in government purchases would not generate movements in current tax rates relative to expected future rates. This conclusion means that, first, variations in G with $\bar{G}$ held fixed have no effect on tax rates; second, shifts in $\bar{G}$ imply equal changes in current and expected future tax rates, $\bar{\tau}$;[17] and third, for the purpose of studying government purchases, it is unnecessary to deal with the intertemporal substitution effects that would arise from expected time variations in income tax rates.[18]

Given this framework for tax rate determination, it is unnecessary to modify the main conclusions that were derived earlier for the case

[13] The analysis neglects the double taxation of incomes that flow through the corporate sector. Taxation in the form of inflationary finance could be included separately without affecting the main results.

[14] Real interest payments on an initial public debt stock would also enter.

[15] Because changes in tax rates affect the tax base, Y, the solution for $\bar{\tau}$ is generally nonunique. However, the minimal possible value for $\bar{\tau}$ is the pertinent choice.

[16] It may be optimal to allow tax rates to vary over the business cycle. A countercyclical pattern shows up empirically for the U.S. federal government. However, the pattern is at least less pronounced in terms of the total government sector.

[17] It is assumed that increases in tax rates induce increases in real tax revenues within the relevant range.

[18] E.g., these types of effects are central to a study of the investment tax credit (see Kydland and Prescott 1977, pp. 482–86).

of a temporary increase in government purchases. Notably, output and the real rate of interest continue to rise when G increases, with $\bar{G}$ held fixed. The principal modification to the previous findings is that r should be interpreted as the after-tax real rate of return, as calculated with the appropriate average marginal tax rate on interest income.[19]

For the case of a permanent increase in government purchases, the new element is the rise in the income tax rate, $\bar{\tau}$, along with the rise in $\bar{G}$ (see n. 17 above). The higher income tax rate motivates a shift away from market work and toward leisure and other nonmarket activities. In equation (3) this change is reflected as downward shifts in the Y^s and C^d functions.[20] For given values of r (now interpreted net of tax), G, $\bar{G}$, and so on, it is plausible that the declines in C^d and Y^s would roughly balance. In other words, as in some cases that were discussed earlier, there is no reason to expect a particular direction of change for the relative amounts of consumption and leisure expenditures that are conducted at different dates.

Given this pattern of response to a higher income tax rate, it still follows that a permanent shift in government purchases has no effect on the (after-tax) real rate of return. However, the negative effect of higher income taxation on the incentive to work offsets the tendency for output to rise.[21] The net movement in output now involves three forces: first, the substitution away from work because of higher income tax rates; second, the negative income effect (associated with the higher level of $\bar{G}$) on leisure, which motivates more work; and third, the direct productive-input effect of G on Y^s. The first two influences involve the standard ambiguous net response of leisure to the substitution and income effects that are generated by either a change in the real wage rate or a shift in the income tax rate that is applicable to labor income. However, the income effect here involves the term $-(1 - \theta - \mathrm{MPG})\bar{G}$, while the tax shift applies one-to-one to $\bar{G}$. This difference increases the likelihood that the substitution effect will outweigh the income effect. Suppose that the substitution effect were, in fact, comparable to or dominant over the income effect. In this case the overall change in output that is induced by a permanent rise in

[19] This result assumes that interest income is taxable and interest payments are deductible from taxable income. The conclusion neglects systematic differences between the marginal tax rates applicable to receivers of interest income vs. those pertinent for payers. The result is not affected directly by the taxation of nominal, rather than real, interest payments. However, other effects of inflation on effective tax rates would matter.

[20] The incentive to accumulate capital would also be diminished. Through this channel, a permanent increase in government purchases tends to reduce the capital stock, even when the after-tax rate of return is unchanged.

[21] A reduction in the capital stock, as mentioned in n. 20, reinforces this effect.

government purchases would be bounded from above by the direct positive effect of G on Y^s. This channel corresponds to only one portion of the positive output response that arose for the case of a temporary increase in government purchases—the other part involves the intertemporal substitution effect on work effort, which is associated with the increase in r. It follows that temporary rises in government purchases would induce larger output responses than equal size, but permanent, rises in purchases. The sign of the output response is now ambiguous for the case of a permanent expansion of government purchases—the reaction is more likely to be positive when the marginal product of public services is high.

Overall, temporary expansions of government purchases are distinguished from permanent increases in that, first, the positive effect on output of the temporary expansion is likely to be larger and, second, a positive effect on the (after-tax) real rate of interest is predicted only for the temporary case. The present empirical investigation deals only with output effects of temporary versus permanent movements in government purchases. Some preliminary analysis of real-rate-of-return effects is carried out in Barro (1981*a*).

II. Empirical Implementation

The theoretical propositions will be tested by examining the effects of government purchases in a reduced-form relationship for output, as measured by real GNP. The analysis is an extension of previous empirical research (Barro 1981*b*), which stressed the business cycle influences of monetary disturbances. This earlier work included a government purchases variable, but did not distinguish temporary from permanent government spending.

It is convenient to carry out the analysis in terms of the ratio of real government purchases to real GNP, G/Y. In particular, temporary or permanent variations in G are assumed to enter relative to Y in a linear relation for the log of output—that is,

$$\log(Y_t) = \ldots + \beta_1[(G - \bar{G})/Y]_t + \beta_2(\bar{G}/Y)_t, \tag{4}$$

where omitted variables indicated by . . . include current and lagged monetary shocks and other deterministic and stochastic influences on output. The variable $\bar{G}_t$, which would generally be unobservable, is the permanent flow of government purchases as perceived at date t. The empirical procedure for handling this variable is discussed below. Lagged values of G/Y and $\bar{G}/Y$ might also appear in equation (4), but these effects were not found to be important empirically.

The functional form in equation (4) implies that increments in

government purchases induce increments in output in accordance with

$$dY[1 + \beta_1(G - \bar{G})/Y + \beta_2\bar{G}/Y] = \beta_1 d(G - \bar{G}) + \beta_2 d\bar{G}. \quad (5)$$

Therefore, in regions where $[\beta_1(G - \bar{G})/Y + \beta_2\bar{G}/Y]$ is small relative to one, the coefficients β_1 and β_2 would indicate the approximate response of Y to unit changes in $(G - \bar{G})$ and $\bar{G}$, respectively. The empirical results suggest that this approximation is satisfactory over the 1930–78 period in the United States, except for the World War II years.

The theoretical analysis suggests testing the hypothesis, $0 \leq \beta_2 \leq \beta_1 \leq 1$. This restriction implies that temporary changes in government purchases have a larger output effect than permanent changes, but that permanent changes also raise (measured) real GNP. These conditions are likely, but not inevitable, implications of the theoretical model.

A central aspect of the empirical analysis is the representation for permanent real purchases as perceived at date t, $\bar{G}_t$. Suppose for the moment that the time-series behavior of G_t implies a relationship for $(\bar{G}/Y)_t$ in terms of a set of parameters α and a vector of currently observed variables, Z_t:

$$(\bar{G}/Y)_t = F(Z_t; \alpha). \quad (6)$$

In this case the unobservable construct, $\bar{G}_t$, could be substituted out from equation (4) to yield a relation for output in terms of observable variables and the vector of unknown coefficients $(\alpha, \beta_1, \beta_2)$,

$$\log(Y_t) = \ldots + \beta_1[(G/Y)_t - F(Z_t; \alpha)] + \beta_2 F(Z_t; \alpha). \quad (7)$$

Some hypotheses arise that concern the role of the Z_t variables in equations (4), (6), and (7). If these variables can be guaranteed, ex ante, not to appear separately in the list of omitted elements that are denoted by . . . in equation (4), then the Z_t variables would appear in equation (7) only to the extent that they serve as determinants for $(\bar{G}/Y)_t$ in the F-function of equation (6). Some cross-equation restrictions therefore emerge for the parameters of equations (6) and (7).

The next sections deal with the problem of modeling a form of equation (6) for real government purchases in the United States.[22]

[22] Levis Kochin has suggested the attractive alternative of using the current overall tax rate as a proxy for the anticipated, long-run average ratio of government purchases to GNP. The rationale for identifying the current tax rate with the anticipated government expenditure ratio was discussed in the theoretical section. Some problems with implementing Kochin's suggestion are: First, the distinction between purchases and expenditures implies that a separate model would be required to predict future transfers (including interest payments), which is not obviously easier than modeling pur-

Government Purchases Equation

The stress on transitory movements in government purchases suggests special attention to war-related expenditures, which are likely to be viewed as largely temporary. I have proceeded empirically by separating total government (federal plus state and local) real purchases of goods and services into a "defense" component, G^w, and other purchases, G^p. The present analysis does not attempt to classify components of government purchases in accordance either with their relative substitutabilities with private spending, as reflected above in the θ parameter, or with their role as inputs to private production, as measured above by the MPG parameter. Differences between defense and nondefense items with respect to these parameters affect the interpretation of some of the empirical findings. Presumably, defense purchases are characterized by a relatively low value of θ and possibly by a relatively high value of MPG. The former implies a relatively large output effect of temporary defense purchases, while the latter would enhance the output effects of both temporary and permanent defense purchases. The empirical analysis would be sharpened by obtaining a division of nondefense purchases into relatively homogeneous categories with respect to the θ and MPG parameters, but the feasibility of this classification is unclear. Transfer payments have not been included in the analysis.

Defense Purchases

A primary determinant of G^w would be the level of current and anticipated future wartime activity, assuming that at least the timing of wars can be treated as exogenous with respect to expenditure decisions. I have quantified this influence by using a casualty rate measure B_t, which represents battle deaths per 1,000 total population (see table 1) for the wartime years since the Civil War: 1898, 1917–18, 1941–45, 1950–53, 1964–72. In effect, this variable can be viewed as an alternative to a set of wartime dummy variables. The casualty rate measure represents an attempt homogeneously to quantify the intensities of different wars and different years within each war, without using military expenditures or personnel measures, which are the types of variables that are to be explained. In particular, the use of

chases directly; and, second, the use of the tax rate to proxy the permanent expenditure ratio may work better for the federal government than for total government. See Benjamin and Kochin (1978), who argue that mobility possibilities would prevent state and local governments from choosing an excess-burden-minimizing debt policy. However, this issue involves also the federal government's interaction with state and local governments—that is, the federal government may compensate for public debt/tax variations that cannot be carried out at the state and local levels.

TABLE 1

CASUALTY RATE VARIABLE

Date	*B*	Date	*B*	Date	*B*
1898	.0052	1945	.603	1966	.025
1917	.23*	1950	.071	1967	.047
1918	.28*	1951	.097	1968	.073
1941	.0044	1952	.030	1969	.046
1942	.162	1953	.021	1970	.021
1943	.205	1964	.0014	1971	.0067
1944	1.090	1965	.0070	1972	.0014

SOURCES.—Vietnam (1964–72). *Statistical Abstract of the United States*, 1977, p. 369, table 590. World War I (1917–18) and Spanish American War (1898): *Historical Statistics of the United States*, 1975, p 1140, line 880 Korean War (1950–53) relative yearly data from Department of the Army, *Battle Casualties of the Army* (1954), were applied to war total from *Statistical Abstract of the United States*, 1977, p 369, table 589 World War II (1941–45) relative yearly data from Office of the Comptroller of the Army, *Army Battle Casualties and Nonbattle Deaths in World War II Final report, December 1941–December 1946*, were applied to war total from *Statistical Abstract of the United States*, 1977, p. 369, table 589 Korean War and World War II data were obtained from William Strobridge, Chief, Historical Services Division, Department of the Army

NOTE.—*B* is battle deaths per 1,000 total population Values of zero apply to dates not listed. Orders-of-magnitude values of *B* (per year) for earlier wars are Revolution (1775–83), 0 2, War of 1812 (1812–15), 0 08, Mexican War (1846–48), 0 04, Civil War (1861–65, union only), 1 0 Casualty figures are from Civil War *Historical Statistics of the United States*, 1975, p. 1140, line 880; other wars. Department of the Army, *History of Military Mobilization in the United States Army*, 1955, appendix A

* Yearly data were unavailable Figures are based on war total assuming equal rate of casualties per month

separate dummy variables for each war would remove any power from the statistical tests that are carried out below.

Because of improvements in the technology of caring for wounded and offsetting changes associated with the "efficiency" of weapons, it is possible that the casualty rate variable does not consistently measure the intensity of war at different dates. I considered using a broader casualty measure that included wounded, but the ratio of this concept to battle deaths showed no trend at least since the Spanish American War.[23] Since I was unable to obtain reliable annual data on wounded for World War II, I have restricted my analysis to the narrower battle deaths concept of casualty rates.

Prospective wars would be likely also to influence current spending, with good information on forthcoming military actions existing prior to at least the U.S. entrances into World Wars I and II. Since I have been unable to construct any instruments for these war expectations, I have introduced some actual future values of *B* into an equation for current defense spending. This procedure introduces errors-in-variables problems into coefficient estimation, although the present analysis is concerned primarily with obtaining conditional forecasts rather than with coefficient estimation, per se. A later part of the

[23] The ratio of total casualties (including wounded, but excluding deaths that were unrelated to combat) to battle deaths is 5.3 for the Spanish American War, 4.8 for World War I, 3.3 for World War II, 4.0 for the Korean War, and 4.3 for the Vietnamese War. See the notes to table 1 for sources of casualty data.

empirical analysis considers a specification where future values of the B variables are excluded from the estimation. Lagged effects of B on spending are introduced also into the equation for defense purchases. Empirically, two annual leads, the contemporaneous value, and up to a third lag of the B variable were found to be important.

Since defense expenditures involve a substantial investment component, the amount of current spending would tend to be influenced negatively by the size of existing capital stocks. Accordingly, I have included in a defense spending equation the variable K^w_{t-1}, which measures the beginning-of-period real stock of military equipment, structures, and inventories (table 2, col. K^w). The relation of capital stock to current spending is assumed to be given by

$$K^w_t = bG^w_t + (1 - \delta)K^w_{t-1}, \tag{8}$$

where δ is a depreciation rate and b measures the fraction of total defense spending that constitutes investment (net of within-year depreciation on this investment). The K^w series was constructed with values of b and δ that varied over time (see the nn. to table 2), but I have limited the theoretical discussion below to situations where these parameters are approximated satisfactorily as constants.

The estimating equation for G^w takes the form

$$g^w_t \equiv (G^w/Y)_t = \alpha_0 B_t + \ldots + \alpha_3 B_{t-3} + a_1 B_{t+1} + a_2 B_{t+2} - \gamma k^w_{t-1} + u_t, \tag{9}$$

where G^w_t is real defense purchases, Y_t is real GNP, $k^w_{t-1} \equiv K^w_{t-1}/Y_{t-1}$, and u_t is a stochastic term. Note that the dependent variable in equation (9) is g^w_t, the ratio of real defense purchases to real GNP. The main part of the subsequent analysis is carried out in terms of ratio variables of this type. The form of equation (9) implies that a doubling of Y_t, K^w_{t-1}, and Y_{t-1}, for given values of the B variables, leads to a doubling of G^w_t. The model for determining g^w_t over time will be used to determine the currently perceived permanent flow of real purchases when expressed relative to real GNP, $\overline{g^w_t} \equiv (\overline{G^w}/Y)_t$.

The error term in equation (9) was modeled satisfactorily as a random walk, so that estimation can be carried out readily in first-difference form:

$$Dg^w_t = \alpha_0 DB_t + \ldots + \alpha_3 DB_{t-3} + a_1 DB_{t+1} + a_2 DB_{t+2} - \gamma Dk^w_{t-1} + \epsilon_t, \tag{10}$$

where D is the first-difference operator and $\epsilon_t \equiv u_t - u_{t-1}$ is a white-noise error term. A constant is insignificant when added to equation (10) in the empirical analysis—that is, there is no trend in the defense purchases ratio. Moving-average error terms or more complicated autoregressive error structures also did not add to the explanatory value of the equation.

The form of equation (10) implies that a current shock ϵ_t—which is not associated with wartime in the sense that the values of the DB variables are small—would have a permanent effect on the future mean level of g^w. Because of the inclusion of the k^w_{t-1} term with a negative sign in equation (10), the effect on $\overline{g^w_t}$ of ϵ_t—that is, of the current actual value g^w_t with values of the B variables and k^w_{t-1} held fixed—turns out to be positive, but somewhat less than one-to-one. On the other hand, if the distribution for future values of B is stationary in level form, positive values for DB tend to be followed at later dates by negative values, which lead to decreases in future values of g^w. In other words wars and the accompanying levels of expenditures are modeled as temporary. This mechanism implies that an increase in g^w_t that is accompanied by positive values of DB will have much less effect on $\overline{g^w_t}$ than will the equivalent peacetime change in defense spending. Wartime spending has an appreciable effect on $\overline{g^w_t}$ only to the extent that military expenditures depart from the amount associated typically with the current level of war intensity. The possibility that wars have a systematically important permanent effect on the purchases ratio is ruled out by the form of equation (10). Some alternative specifications of the error process that would have admitted this type of persisting effect were not supported empirically for the United States.

As detailed in the Appendix, equations (10) and (8) can be used to express expected future values of the defense purchases ratio, g^w_{t+i}, as a function of the latest observed ratio g^w_t, the value k^w_{t-1}, and actual and expected future values of the war intensity variable, B. Summation over these expressions with proper allowance for discounting yields a relation for the permanent purchases ratio $\overline{g^w_t}$ in terms of g^w_t, k^w_{t-1}, the array of B variables, and a real discount rate ρ (which equals the difference between the real interest rate and the growth rate of real GNP). Finally, a simple specification for the stochastic structure of the B variable—based on the frequency, intensity, and duration of wars over the full history of the United States—is used to solve for the expected future values of the B variables. With this substitution, $\overline{g^w_t}$ is determined as a function of observable magnitudes, up to the setting of a discount rate. In the case where future values of the B variables are admitted into the government purchases equation (10), $\overline{g^w_t}$ ends up as a function of the values $B_{t+2}, \ldots, B_{t-3}$—that is, the values B_{t+1} and B_{t+2} are treated as observable at date t. In a situation where future values of B are excluded from equation (10), $\overline{g^w_t}$ is expressed in terms of the current and lagged values, $B_t, \ldots, B_{t-3}$ (as well as the values of g^w_t and k^w_{t-1}).

The main product of this exercise from the full empirical analysis is the series for temporary real defense purchases expressed relative to

TABLE 2

GOVERNMENT PURCHASES VARIABLES

Date	g^{w}*	$\overline{g^{w}}$†	$g^{w} - \overline{g^{w}}$	k^{w}‡	g^{p}§
1889	.0060	...	...	...	.097
1890	.0057	...	...	...	.094
1891	.0057	...	...	...	.094
1892	0056	..	.	. .	.089
1893	.0062	...	...	...	.097
1894	.0064	...	...	...	.102
1895	.0053	...	...	...	.092
1896	.0059	...	...	...	.097
1897	.0080	...	...	...	.091
1898	.0192	...	...	...	.094
1899	.0165	...	...	...	.087
1900	.0131	...	...	...	.088
1901	.0113	...	...	...	.081
1902	.0110	...	. .	...	.084
1903	.0108	...	...	..	.088
1904	.0123	. .	...	...	.088
1905	.0112	...	...	...	.089
1906	.0093	..	...	...	.082
1907	0090	...	...	...	.091
1908	0117	...	...	...	.110
1909	.0105	.	...	...	.089
1910	.0100	...	..	...	091
1911	0101	..	...	...	.105
1912	.0094	...	..	...	.100
1913	0096	..	...	...	.096
1914	0139	..	...	...	.106
1915	0135	...	...	..	.112
1916	.0164	...	..	...	.093
1917	.076	...	...	...	.085
1918	258	...	...	...	080
1919	.156	...	...	...	.049
1920	038	...	..	..	.085
1921	.033	...	...	...	.125
1922	.017	..	...	..	.116
1923	014	...	...	...	.105
1924	.014	...	...	...	.116
1925	.012	...	...	...	.115
1926	.011	...	...	...	.108
1927	.012	...	...	...	.118
1928	013	...	...	...	.121
1929	.013	...	..	.055	.117
1930	.015	.030	−.015	056	.141
1931	.017	.031	− 014	.057	.158
1932	.019	.033	−.014	.063	.175
1933	.016	.032	−.016	.063	.176
1934	.016	.032	−.016	.056	.187
1935	.017	.031	−.014	.050	.174
1936	.018	.031	−.013	.045	.179
1937	.016	.029	−.013	.043	.165
1938	019	.031	− 012	.046	.187
1939	.017	.029	−.012	.045	.183
1940	.028	.025	.003	.052	.163

TABLE 2 (*Continued*)

Date	g^{w}*	$\overline{g^{w}}$†	$g^{w} - \overline{g^{w}}$	k^{w}‡	g^{p}§
1941	.120	.065	055	074	.122
1942	.317	.104	.213	.164	.092
1943	.439	.093	.346	.309	.069
1944	463	.074	389	.431	.064
1945	.410	.069	.341	.470	.063
1946	.103	063	.040	.483	.092
1947	.055	.069	−.014	.414	.106
1948	.056	.072	−.017	337	.117
1949	.064	.091	−.026	.290	.132
1950	.066	.074	−.008	.239	.117
1951	.123	.098	.026	209	.107
1952	156	.127	.030	225	.110
1953	.156	.140	.016	260	.118
1954	.133	132	.001	.291	.119
1955	.115	132	−.017	.284	.116
1956	.112	:129	− 017	283	.115
1957	.116	.133	−.017	.286	119
1958	.115	.133	−.018	.296	134
1959	108	.129	−.022	.286	129
1960	.102	.124	−.022	282	.133
1961	.104	.124	− 021	.278	.138
1962	103	.123	−.020	.268	.138
1963	.096	.116	− 021	263	142
1964	.087	.107	−.021	251	.145
1965	.080	.096	−.016	.235	.147
1966	088	.090	−.002	.219	146
1967	098	.086	.011	216	149
1968	096	081	.014	.211	.151
1969	.088	.080	.009	.206	.150
1970	.079	.083	−.004	201	154
1971	.068	.080	− 012	173	.157
1972	.063	076	− 014	169	.153
1973	056	.074	− 017	.153	.148
1974	.055	071	−.017	147	157
1975	.055	070	−.016	.141	.164
1976	.050	.067	− 016	128	.156
1977	.049	.064	− 015	.118	152
1978	.046	.060	− 014	.111	.150

* $g^{w} = G^{w}/Y$, where Y is real GNP (1972 base) G^{w} is real defense purchases (1972 base) Data since 1929 are from *National Income and Product Accounts of the United States* and recent issues of the *United States Survey of Current Business* The fraction of nominal defense purchases in total nominal federal purchases was multiplied by figures on real federal purchases (1972 base) Data from 1889–1928 are from Kendrick (1961, table A-1, col 5) Figures were multiplied by 4 8, based on the overlap for 1929.

† $\overline{G^{w}}$ is the estimated normal defense purchases ratio, as calculated from eq (14) in the text

‡ $k^{w} = K^{w}/Y$, where K^{w} is the end-of-year value of net real stocks of military structures, equipment, and inventories (1972 base) Data from 1929–69 are from Kendrick (1976, table B-24) converted from a 1958 to a 1972 index by a constant multiple (1 72) Figures were extended to 1978 using data on various expenditure components: military structures, AEC structures, military equipment, AEC equipment, inventories for GSA stockpiles, and inventories for AEC stockpiles Depreciation estimates were based on rates used by Kendrick within each category His calculations assume a higher rate of depreciation during World War II

§ $g^{p} = G^{p}/Y$, where G^{p} is real nondefense purchases of the federal plus state and local government sectors (1972 base). G^{p} was calculated as total real government purchases G less G^{w} Sources for G correspond to those above for G^{w}, except that Kendrick (1961, table A-IIa) was used for data from 1889–1928

real GNP, $(g^w - \overline{g^w})_t$, which is indicated over the 1930–78 period in table 2. (This series corresponds to the choice of discount rate—the difference between the real rate of interest and the growth rate of real GNP—of .02 per year.) As is evident from the table, this variable identifies the years associated with wartime, 1940–46, 1951–53, 1966–69, as times when the defense purchases ratio is above its perceived long-run average value. Although the underlying model allows a quantitative assessment of the gap between g_t^w and $\overline{g_t^w}$ in the context of wars with different intensities and in an environment of nonconstant values for $\overline{g_t^w}$, it is also clear that the general pattern for the $(g^w - \overline{g^w})_t$ variable would be robust to some changes in the underlying model. Notably, the important aspect of the stochastic specification for the war intensity variable B is the temporary nature of wars, rather than the details of war probabilities. Substantial variations in the discount rate ρ also have minor effects on the results.

It should not be surprising that the $(g^w - \overline{g^w})_t$ variable exhibits a substantial amount of positive serial correlation. In this respect a gap between current and normal values—which the $(g^w - \overline{g^w})_t$ variable is intended to capture—should be distinguished from the spread between actual and anticipated or perceived amounts, which has been stressed in earlier analyses of monetary disturbances (Barro 1981*b*). The latter type of variable exhibits serial independence as a consequence of rational expectations and the assumption that information is received with, at most, a one-period lag. This type of argument does not apply to a variable that measures temporary effects. In the case of the temporary defense purchases variable, the large number of (serially correlated) peacetime years with small negative values of $(g^w - \overline{g^w})_t$ is offset by a smaller number of (serially correlated) wartime years with excesses of g_t^w over $\overline{g_t^w}$. (However, the years that are significantly affected by war—for example, the set 1941–46, 1950–53, 1965–72—should not be deemed special, since they constitute 47 percent of the years since 1941 and 39 percent of those since 1946.)

Government Purchases of Nondefense Items

The nondefense portion of government purchases—$g^p \equiv G^p/Y$, where G^p is nondefense real purchases—was examined statistically over samples beginning in 1929. This study revealed little predictive value for first differences, Dg^p, except for a negative association with the contemporaneous change in the defense component, Dg^w.[24] In par-

[24] Past history of the residuals, lagged values of Dg^p or Dg^w, a capital stock measure Dk^p, and a constant were all insignificant.

ticular, there is no drift in the nondefense purchases ratio. The negative association of Dg^p with Dg^w reflects the crowding out of nondefense government spending during wartime. The dependence of Dg^p only on Dg^w means that departures of g^p from the perceived normal value $\overline{g^p}$ are determined completely by the difference between g^w and $\overline{g^w}$. With $g^w - \overline{g^w}$ held fixed, changes in g^p amount entirely to shifts in the permanent component of nondefense purchases. Accordingly, with the g^w variables entered separately, the coefficient of the g^p variable in an output equation would reveal the effects of permanent changes in nondefense purchases. It is not possible here to estimate the response of output to temporary changes in nondefense purchases, since no temporary changes were isolated over the sample.

Empirical Results

The principal empirical analysis involves joint estimation of the government purchases equation (10) and a relation for output that is based on the form of equation (4). With the defense and nondefense components of government purchases entered separately, the output equation becomes

$$\log(Y_t) = \ldots + \beta_1(g^w - \overline{g^w})_t + \beta_2\overline{g_t^w} + \beta_3 g_t^p. \qquad (11)$$

Note that the real government purchases variables all appear as ratios to real GNP. The variable $\overline{g_t^w} \equiv (\overline{G^w}/Y)_t$ is determined as a function of observables from equation (10) when used in conjunction with some relations that are derived in the Appendix (eqq. [A5] and [A6]).

The first set of hypothesis tests involves the output effects of the $(g^w - \overline{g^w})_t$ and $\overline{g_t^w}$ variables in equation (11)—specifically, that the coefficients of these variables satisfy the restrictions, $0 \leq \beta_2 \leq \beta_1 \leq 1$. As indicated earlier, the coefficient on the g_t^p variable, β_3, reveals the output effect of a permanent change in nondefense purchases. If nondefense purchases were characterized by closer substitutability with private consumption expenditure (the θ parameter) and by lesser impact on private production (the MPG parameter) than defense purchases, then $\beta_2 > \beta_3$ would follow. However, this condition cannot be viewed as a firm implication of the theory. The model also suggests the restriction, $0 \leq \beta_3 \leq 1$.

The second set of hypothesis tests checks whether the explanatory variables for $\overline{g_t^w}$—in this case $B_{t+2}, \ldots, B_{t-3}$ and k_{t-1}^w—enter an unrestricted reduced form for output as determined solely by their role in determining $\overline{g_t^w}$ in accordance with the coefficients of equation (10).

The analysis is contingent on a value of the discount rate ρ—the difference between the real interest rate and the growth rate of real

GNP—in the calculation of the $\overline{g_t^w}$ variable (see eqq. [A5] and [A6] in the Appendix). However, the results turn out to be relatively insensitive to variations in the ρ parameter at least over the range from .01 to .05 per year. The reported results refer to a fixed value of $\rho = .02$ per year, which is a plausible magnitude, ex ante, and which approximates the maximum likelihood estimate for this parameter.

Jointly estimated equations for real defense purchases and real GNP were calculated by means of a nonlinear, maximum likelihood routine from the TSP regression package, which includes estimation of contemporaneous covariances for the error terms. The estimation is joint in the sense of incorporating the role of the coefficients from equation (10) in determining the series $\overline{g_t^w}$ and thereby influencing the fit for output in the form of equation (11). In particular, the coefficients in the equation for Dg_t^w are not determined solely to obtain a best fit of equation (10). I have not carried out joint estimation in the broader context of choosing the number of leads and lags of the B variable to include equation (10), in deciding to omit moving-average error terms in this equation, in analyzing the process for nondefense purchases, and so on.

Since the dependent variable, real GNP, appears also in the denominators of the ratio variables g^w and g^p, there is a possible simultaneity problem in the estimation. Accordingly, I have used as instruments for g_t^w and g_t^p the lagged values, g_{t-1}^w and g_{t-1}^p, and also the contemporaneous values, $G_t^w/\bar{Y}_t$ and $G_t^p/\bar{Y}_t$, where $\bar{Y}_t$ is the trend value of real GNP, as determined from a regression over the 1946–78 period of log(GNP) on a constant and time. The estimates are not altered substantially if only the pair of lagged values or only the pair of contemporaneous values relative to trend are used as instruments. Empirically, the movements in g_t^w are dominated sufficiently by conditions of war or peace that the use of instruments yields estimates for the coefficients of the $(g^w - \overline{g^w})_t$ and $\overline{g_t^w}$ variables in equation (11) that differ only in minor ways from ordinary-least-squares (OLS) values. However, the use of instruments is important in the case of the g_t^p variable—OLS estimates for the β_3 coefficient in equation (11) appear to be biased downward substantially because of the inclusion of Y_t in the denominator of the g_t^p variable.

Results for Post–World War II Output Sample

For an output sample that begins in 1946, the results of the joint estimation of equations (10) and (11) are, for the 1932–78 sample:

$$\underset{}{Dg_t^w} = \underset{(.013)}{.163DB_{t+2}} + \underset{(.012)}{.198DB_{t+1}} + \underset{(.014)}{.273DB_t} + \underset{(.017)}{.240DB_{t-1}}$$

$$- \underset{(.015)}{.022}DB_{t-2} + \underset{(.016)}{.088}DB_{t-3} - \underset{(.08)}{.26}Dk^{w}_{t-1}, \tag{12}$$

$$\hat{\sigma} = .0143, \text{ D-W} = 1.7;$$

and for the 1946–78 sample:

$$\log(Y_t) = \underset{(.04)}{2.97} + \underset{(.0008)}{.0343} \cdot t + \underset{(.22)}{.83}DMR_t + \underset{(.22)}{1.12}DMR_{t-1}$$

$$+ \underset{(.21)}{.99}(g^{w} - \overline{g^{w}})_t + \underset{(.12)}{.55}\overline{g^{w}_t} + \underset{(.45)}{.62}g^{p}_t, \tag{13}$$

$$\hat{\sigma} = .0143, \text{ D-W} = 1.5.$$

Asymptotic standard errors are shown in parentheses below the coefficient estimates. The $\hat{\sigma}$ values are asymptotic estimates of the standard errors of the disturbance terms; D-W is the Durbin-Watson statistic. Note that the sample for the government purchases equation (12) begins in 1932 and thereby includes the World War II experience.

Variables included in equations (12) and (13) are: Y, real GNP (1972 base); t, time trend; $DMR \equiv DM - \hat{DM}$ is "unanticipated money growth," as measured in earlier research (Barro 1981*b*), where $\hat{DM}$ is an estimated value of money growth from an equation that is based on the $M1$ definition of the money stock;[25] $g^{w} \equiv G^{w}/Y$, where G^{w} is real defense purchases (1972 base); $g^{p} \equiv G^{p}/Y$, where G^{p} is real, nondefense, federal plus state and local purchases (1972 base); B, casualty rate variable as defined in table 1; and $k^{w} \equiv K^{w}/Y$, where K^{w} is real government defense capital stocks (1972 base).

For present purposes I focus on the role of the government purchases variables in equation (13). The money shock variables have effects that are similar to those discussed in previous research, as reported in Barro (1981*b*).

The $\overline{g^{w}_t}$ variable in the output equation is based on the specification for Dg^{w}_t that appears in equation (12). The main result from this equation is the strong positive spending effect of wars, as measured by the casualty rate variable B. The equation shows a 2-year lead effect of the B variable and a lagged effect out to 3 years. (The negative effect on Dg^{w}_t of the DB_{t-2} variable is difficult to interpret.) The consequences of eliminating the future values of DB from this equation are

[25] $\hat{DM}$ is determined from an equation that is estimated over the 1941–78 sample:

$$\hat{DM}_t = \underset{(.024)}{.095} + \underset{(.14)}{.49}DM_{t-1} + \underset{(.12)}{.16}DM_{t-2} + \underset{(.015)}{.069}\text{FEDV}_t + \underset{(.008)}{.030} \cdot \log[U/(1-U)]_{t-1},$$

where observations from 1941–45 are weighted by .36. FEDV_t is real federal spending relative to a distributed lag of itself, and U is the unemployment rate in the total labor force. See Barro (1981*b*) for a discussion of this type of equation.

discussed later. For present purposes the most important aspect of war spending is its temporary nature, although precise calculations for $\overline{g_t^w}$ involve the distributed lag pattern of DB effects on Dg_t^w and the implications of these responses for the behavior of the capital stock ratio k^w. Equation (12) shows also the expected negative effect of Dk_{t-1}^w on Dg_t^w.

Using equations (A5), (A6), and (A11) from the Appendix, and the value $\rho = .02$ for the discount rate, the point estimates of coefficients that appear in equation (12) can be shown to imply the formula for $\overline{g_t^w}$ as follows:

$$\begin{aligned}\overline{g_t^w} = {} & .011 + .67g_t^w + .15k_{t-1}^w - .06B_{t-3} + .02B_{t-2} - .16B_{t-1} \\ & - .18B_t - .12B_{t+1} - .07B_{t+2}.\end{aligned} \tag{14}$$

This relation is a particular form of equation (6). Equation (14) shows a positive but less than one-to-one effect on $\overline{g_t^w}$ of g_t^w, a positive effect of k_{t-1}^w (for a given value of g_t^w), and a basically negative effect of the casualty rate variables (again given the value of g_t^w). Values of $\overline{g_t^w}$ that are calculated from equation (14) are shown along with values of g_t^w in table 2.

The temporary defense purchases variable, $(g^w - \overline{g^w})_t$, has a significantly expansionary effect on output. The estimated coefficient[26] in equation (13) is $\hat{\beta}_1 = .99$, S.E. $= .21$. The "t-value" corresponding to $\beta_1 = 0$ is 4.7. The normal defense purchases variable, $\overline{g_t^w}$, is also significantly expansionary in this equation—$\hat{\beta}_2 = .55$, S.E. $= .12$, which implies a t-value of 4.6. The estimated effect for the permanent purchases variable is somewhat greater than half that of the estimated temporary effect.[27] The results permit rejection of two extreme hypotheses: first, that only the temporary part of purchases affects output (which would require $\hat{\beta}_2$, the estimated coefficient of the $\overline{g_t^w}$ variable in eq. [13], to differ insignificantly from zero),

[26] Because of the negative correlation of $(g^w - \overline{g^w})_t$ with g_t^p, the β_1 coefficient picks up an additional effect. The extra term involves the difference between the output effects of permanent and temporary nondefense purchases. The output coefficient associated with temporary nondefense purchases could not be estimated with the available data. However, since the regression coefficient of g_t^p on $(g^w - \overline{g^w})_t$ is on the order of -0.1, it is unlikely that the overall modification is important.

[27] It has been suggested that the temporary government purchases variable may be proxying for the effects of accompanying federal deficits. The analysis in Barro (1979) documents the strong positive effect of temporary federal spending, as in wartime, on public-debt issue. Some preliminary results in Barro (1980*b*) indicate that lagged "debt shocks" have expansionary effects on output that are statistically significant but substantially weaker than those of monetary shocks. However, this constructed debt-shock variable filters out the normal positive association between temporary government spending and the deficit. With these debt shocks held fixed, the actual lagged values of public-debt growth have no explanatory value for output. This last finding suggests that the strong expansionary influence of temporary defense purchases does not involve a proxying for the effect of correlated movements in the federal deficit.

and second, that temporary and permanent purchases are of equal importance for output. The latter case would correspond to equal coefficients ($\beta_1 = \beta_2$) for the $(g^w - \overline{g^w})_t$ and $\overline{g_t^w}$ variables—that is, to the proposition that the coefficient of the $\overline{g_t^w}$ variable would be zero in an equation that held fixed the value of the actual purchases ratio, g_t^w. For convenience, the results from equation (13) can be rewritten in this form as

$$\log(Y_t) = \ldots + \underset{(.21)}{.99 g_t^w} - \underset{(.24)}{.44 \overline{g_t^w}}.$$

The hypothesis that the coefficient of the $\overline{g_t^w}$ variable equals zero corresponds to a t-value of 1.8, which is significant at the 5 percent level for the case of this one-sided test. That is, the null hypothesis of equal output effects for temporary and permanent defense purchases ($\beta_1 = \beta_2$ in eq. [11]) is rejected in favor of the hypothesis that temporary purchases are more expansionary, $\beta_1 > \beta_2$.[28]

The estimated coefficient on the $(g^w - \overline{g^w})_t$ variable in equation (13) implies that a temporary change in the level of real defense purchases has almost a one-to-one effect on the contemporaneous level of output. While this finding is consistent with the restriction, $\beta_1 \leq 1$, the evidence would also be consistent with a moderate multiplier relationship between temporary government purchases and output. The relatively high estimated output effect is associated in the theoretical model with a small value of the θ-coefficient, a high value of the MPG parameter, and a high real-rate-of-return elasticity of aggregate supply relative to that of demand.

The estimated coefficient on the $\overline{g_t^w}$ variable in equation (13) implies that a permanent increase by one unit in real defense purchases leads approximately to a one-half unit rise in real GNP. This result accords with the restriction $\beta_2 \leq 1$—moreover, the estimate is significantly below unity in this case.

The estimated coefficient of g_t^p, the nondefense purchases ratio, is

[28] I considered discriminating between temporary and permanent defense purchases by utilizing a measure of the return on the equities of defense contractors relative to that on a market portfolio of New York Stock Exchange (NYSE) stocks. (The relative-returns variable was constructed using data on total returns to NYSE stocks from the Center for Research in Security Prices of the University of Chicago. A list of defense contractors and the size of these contracts for 1969 from the Department of Defense was kindly supplied to me by Claire Friedland.) This relative-returns variable has no explanatory power when added to a first-difference form of the output equation with Dg_t^w, Dg_t^p, and the determinants of $D\overline{g_t^w}$ from the form of eq. (12) included as independent variables. Conceptually, it is unclear whether the relative-returns variable signifies an increase in war probability and, therefore, that current defense expenditures are more likely to be temporary, or an increase in the long-run expected quantity of defense purchases, which would imply that current defense expenditures are more likely to be permanent. Therefore, the sign of the variable is ambiguous on theoretical grounds.

$\hat{\beta}_3 = .62$, S.E. $= .45$, which is positive but imprecisely determined. The relatively small amount of independent sampling variation in this variable since 1946, in conjunction with the necessity of using instrumental variables, results in the high standard error. (Using OLS techniques, the estimated coefficient of the g_t^p variable is negative.) In any event the hypothesis that the coefficients of the g_t^p and $\overline{g_t^w}$ variables are equal, $\beta_2 = \beta_3$, would be accepted from the present evidence. As noted earlier, the relative output effects for these two types of permanent movements in government purchases depend, in the theoretical model, on the relative values of the θ and MPG parameters. Possibly, a more precise determination of the g_t^p coefficient would be obtained by extending the analysis to the 1930s, during which major changes occurred in the nondefense purchases ratio. The main obstacle for this extension is the isolation of monetary shocks, which seems to require a specification for the pre–World War II monetary regime that differs from that used for the post–1941 period (see n. 25 above).[29]

A combination of equation (11) with an expression for $\overline{g_t^w}$ in the form of equation (14) implies a reduced-form relation for output in terms of a constant, a time trend, *DMR* variables, g_t^w, g_t^p, the *B* variables, and k_{t-1}^w. Unrestricted estimation of this reduced form affords a test of the hypothesis that the determinants of $\overline{g_t^w}$—specifically, the *B* variables and k_{t-1}^w—affect output only in the manner implied by the forms of equations (11) and (14). The test is based on the likelihood ratio corresponding to unrestricted and restricted forms of joint estimation. The value of $-2 \cdot \log(\text{likelihood ratio})$ turns out to be 3.2, which is below the 5 percent critical value for the χ^2 distribution with 6 degrees of freedom (the number of coefficient restrictions in this case) of 12.6.[30] Therefore, the hypothesis that the determinants of $\overline{g_t^w}$ enter only in this indirect manner in influencing output is accepted.

[29] It is worth noting that the point estimates for the government purchases coefficients in eq. (13) change little if the output equation is respecified in first-difference form (see Plosser and Schwert [1978] for a discussion). The estimated equation in this case is, for the 1946–78 sample:

$$D\log(Y_t) = \underset{(.006)}{.036} + \underset{(.29)}{.83}D(DMR_t) + \underset{(.34)}{1.01}D(DMR_{t-1}) + \underset{(.22)}{.91}D(g^w - \overline{g^w})_t$$
$$+ \underset{(.39)}{.55}D\overline{g_t^w} + \underset{(1.67)}{0.63}Dg_t^p, \quad \hat{\sigma} = .0176, \text{D-W} = 2.5.$$

The jointly estimated equation for Dg_t^w is similar to that shown in eq. (12). Note that the constant in the equation for $D\log(Y_t)$ corresponds to the time trend for the level equation. The main change from the previous specification is the higher standard errors for the $\overline{g_t^w}$ and g_t^p coefficients. The Durbin-Watson statistic of 2.5 suggests overdifferencing.

[30] If the discount rate ρ were regarded as a freely estimated parameter, there would be only 5 degrees of freedom, which would imply a critical value of 11.1.

There would, of course, be many possible output effects of war that do not operate through the channels that were specified in the present model. For example, there would be responses to conscription and patriotism, and the possibility that war would threaten future property rights. Some of these effects would, however, influence output in a manner similar to that of the pecuniary intertemporal substitution variable that was stressed in the theoretical analysis (see Barro [1981*a*] for a discussion of wartime influences on financial rates of return). In particular, the level of temporarily high demands on resources by the government, $(g^w - \overline{g^w})_t$, may proxy satisfactorily for the full range of wartime output effects.

More generally, because the war variables are the prime basis in this work for distinguishing temporary from permanent movements in government purchases, it would be infeasible to allow unrestricted direct wartime output effects and still carry out interesting tests of the underlying hypotheses. In any event the restriction that the war variables influence output only indirectly through influences on temporary government purchases is satisfied in the present case.

Elimination of future values of the casualty rate variable from the government purchases equation (12) has a substantial effect on the estimation of this equation for the 1932–78 sample. Aside from a major deterioration in fit, the residuals then show pronounced positive serial correlation. These effects are dominated by the World War II years—in particular, from the rise in military spending in 1940–41 prior to the onset of casualties and from the major advance in spending in 1942–43 before the peak in casualties for 1944–45. In these cases it seems reasonable to treat the future casualty values as rough proxies for contemporaneously available information about the intensity of the war. (Another possibility would be to use a foreign casualty rate variable, but data limitations and conceptual problems concerning the perceived threat attached to foreign conflicts have prevented the implementation of this idea.)

For the post–World War II period, this type of advance information on war intensity seems less important and, in fact, the future casualty variables lack significant explanatory power for defense expenditures over this period. With these future values deleted and 1946–78 samples used throughout, the results of the joint estimation are, for the 1946–78 sample:

$$\begin{aligned} Dg_t^w = {} & \underset{(.061)}{.264}DB_t + \underset{(.043)}{.226}DB_{t-1} - \underset{(.021)}{.027}DB_{t-2} + \underset{(.022)}{.079}DB_{t-3} \\ & - \underset{(.11)}{.30}Dk_{t-1}^w, \qquad \hat{\sigma} = .0128, \text{D-W} = 1.9; \end{aligned} \tag{15}$$

and

$$\log(Y_t) = \underset{(.06)}{2.99} + \underset{(.0010)}{.0340} \cdot t + \underset{(.25)}{.88}DMR_t + \underset{(.26)}{1.12}DMR_{t-1}$$
$$+ \underset{(.26)}{.96}(g^w - \overline{g^w})_t + \underset{(.17)}{.52}\overline{g_t^w} + \underset{(.52)}{.74}g_t^p, \tag{16}$$
$$\hat{\sigma} = .0177, \text{ D-W} = 1.5.$$

Although the fit of the estimated output equation (16) is poorer than that shown in equation (13), the pattern of estimated coefficients and standard errors is similar. Therefore, the conclusions on output effects of the government purchases variables are not sensitive to, first, elimination of the World War II years from the sample for the government purchases equation and, second, removal of the future values of the casualty rate variable from this equation. The results also remain similar if the starting date for the output equation is shifted from 1946 to 1950, which removes some contribution of the World War II years that works through the effects of lagged B values on the constructed $\overline{g_t^w}$ variable.

Addition of World War II Output Experience

Rather than insulating the results from World War II, it is in many respects more informative to evaluate the performance of the model during this extreme experience. Clearly, the sample variation in the wartime-related variable, $(g^w - \overline{g^w})_t$, is raised enormously by this extension of coverage. On the other side, the inclusion of the World War II years raises problems that concern the measurement of real output during a period of extensive price controls and the accuracy of linear specifications for extreme observations.

Jointly estimated equations that include the 1942–45 observations on output are, for the 1932–78 sample:

$$Dg_t^w = \underset{(.013)}{.163}DB_{t+2} + \underset{(.013)}{.196}DB_{t+1} + \underset{(.015)}{.274}DB_t + \underset{(.017)}{.242}DB_{t-1}$$
$$- \underset{(.015)}{.022}DB_{t-2} + \underset{(.016)}{.087}DB_{t-3} - \underset{(.08)}{.25}Dk_{t-1}^w, \tag{17}$$
$$\hat{\sigma} = .0143, \text{ D-W} = 1.7;$$

and for the 1942–78 sample:

$$\log(Y_t) = \underset{(.05)}{2.97} + \underset{(.0009)}{.0350} \cdot t + \underset{(.15)}{.76}DMR_t + \underset{(.15)}{.90}DMR_{t-1} \tag{18}$$

$$+ \underset{(.06)}{.71}(g^w - \overline{g^w})_t + \underset{(.12)}{.42}\overline{g_t^w} + \underset{(.51)}{.14}g_t^p,$$

$$\hat{\sigma} = .0154, \text{ D-W} = 1.4.$$

A test that the 1942–45 observations for output conform to the same structure as that for the other years corresponds to a value for $-2 \cdot \log(\text{likelihood ratio})$ of 37.9, which exceeds the 5 percent critical value of the χ^2 distribution with 4 degrees of freedom of 9.5.[31] It seems likely that this appearance of structural break during World War II would not appear if the functional form were altered to allow for a nonlinear dependence of $\log(Y_t)$ on $(g^w - \overline{g^w})_t$. Specifically, the most important change from the 1946–78 estimates in equation (13) to the 1942–78 values in equation (18) seems to be the drop in the estimated coefficient of the $(g^w - \overline{g^w})_t$ variable, $\hat{\beta}_1$, from .99, S.E. = .21, to .71, S.E. = .06. A functional representation that allowed for positive, but diminishing, output effects of temporary government purchases would probably account for the overall results in a homogeneous form, but I have not experimented along these lines. As the results in equation (18) stand, they reveal the anticipated reduction in the standard error of the $(g^w - \overline{g^w})_t$ coefficient—in fact, if this relation were viewed as well specified, the estimated output effect of temporary government purchases, $\hat{\beta}_1$, would now be measured as significantly less than one.[32]

[31] I have also carried out the estimation with an allowance for heteroscedasticity in the form of a different error variance for the output equation during the World War II years, 1942–45. Maximum likelihood estimates indicate that those years have an error variance that is 2.6 times that for the 1946–78 period, which corresponds to multiplying the 1942–45 observations by .62 in the estimation. The results with the heteroscedasticity correction applied are, for the 1942–78 sample.

$$\log(Y_t) = \underset{(.04)}{2.98} + \underset{(.0009)}{.0347} \cdot t + \underset{(.17)}{.85}DMR_t + \underset{(.17)}{1.01}DMR_{t-1}$$

$$+ \underset{(.07)}{.75}(g^w - \overline{g^w})_t + \underset{(.12)}{.48}\overline{g_t^w} + \underset{(.53)}{.34}g_t^p, \ \hat{\sigma} = .0148, \text{ D-W} = 1.4$$

Note that the $\hat{\sigma}$ value applies here to the error term for the 1946–78 period. The pattern of results does not differ greatly from that shown in eq. (18). The jointly estimated government purchases equation is very close to that shown in eq. (17). A test for the addition of the 1942–45 years for output to the rest of the sample corresponds here to a value for $-2 \cdot \log(\text{likelihood ratio})$ of 30.9, as compared to a 5 percent χ^2 value with 3 degrees of freedom of 7.8. (The degrees of freedom are reduced by 1 here in comparison with the test for the context of homoscedasticity because of the estimation of the heteroscedasticity parameter.)

[32] The estimated coefficient for nondefense purchases, g_t^p, is also smaller than before, but still insignificantly different from that on the $\overline{g_t^w}$ variable. With g_t^p held fixed, the implied estimated coefficient on $\overline{g_t^w}$ is now .29, S.E. = .12. That is, the hypothesis that temporary and permanent defense purchases have equal output effects—$\beta_1 = \beta_2$—corresponds here to a t-value of 2.4. This hypothesis would, therefore, again be rejected in favor of $\beta_1 > \beta_2$. A test that the determinants of $\overline{g_t^w}$ from eq. (17) enter only in this

Expectations of Future Wars

Calculation of expected future casualty rates is based on the following stationary probability model for wars.[33] First, a 2 × 2 matrix is specified for the probability of war or peace next year (or rather for year $t + 3$ when conditions at $t + 2$ are assumed known at date t), conditional on war or peace prevailing currently. It is assumed that information about the future course of B is contained fully in the most recent observation, earlier values of B and values of other variables not having to be considered. The probability of war during at least part of next year, given peace for the latest observation, is based on data over the 1774–1978 period—namely,

$$p_1 = \text{Prob}(B_{t+1} > 0 \mid B_t = 0) = 9/162 = .06, \tag{A7}$$

where 162 is the total number of peacetime years in the sample (where $B_t = 0$), and 9 is the number of these years that were followed by the outbreak of war.[34] Correspondingly, the probability of peace continuing is given by

$$(1 - p_1) = \text{Prob}(B_{t+1} = 0 \mid B_t = 0) = .94.$$

The value of p_1 is slightly higher if the sample is limited to the more recent period 1889–1978 (the sample for which relatively accurate observations on G^w and B are available), for which the result is $p_1 = 5/68 = .07$.

The probability of the continuation of war is given for the 1774–1978 sample by

$$p_2 = \text{Prob}(B_{t+1} > 0 \mid B_t > 0) = 33/42 = .79, \tag{A8}$$

where 42 is the number of war years (where $B_t \neq 0$), and 33 is the number of these that were followed by another year of war.[35] In other words nine wars began and ended over the sample 1774–1978. For the 1889–1978 period, the result would be $p_2 = 16/21 = .76$. Finally, the probability of no war next year, given its existence this year,[36] is given for the 1774–1978 sample by

$$(1 - p_2) = \text{Prob}(B_{t+1} = 0 \mid B_t > 0) = .21.$$

The expected value of B for the first year of a war is calculated as the mean value for the five wars since 1889 (for which accurate data on B are available):

[33] War probabilities and the distribution of sizes of wars need not be constant over time, although there is no indication of substantial structural change in the small sample of evidence afforded by the 200 years of U.S. history. (The largest value for the B variable would actually apply to the Civil War—see the notes to table 1.) From the standpoint of constructing the $\overline{g^w}$ variable, shifts in the stochastic structure for wars would essentially be an alternative to the present specification that allows for shifts in spending for a given war structure, as represented by the stochastic variable ϵ_t in eq. (A1). In the context of output analysis, it is unclear that there would be much empirical difference between these alternatives.

[34] The year 1978 is not included in this calculation, although it could have been if peace during 1979 were also included. War years are taken to be 1775–83, 1812–15, 1846–48, 1861–65, 1898, 1917–18, 1941–45, 1950–53, and 1964–72. There may be some objection to starting the sample just before a war (which is not independent of the start of U.S. data), but the results are not highly sensitive to this choice.

[35] The probability p_2 refers to the existence of war during at least part of a year following a period of war during at least part of the previous year.

[36] This calculation pertains to the existence of peace over the entire year $t + 1$, conditional on war during at least part of year t.

TABLE A1

PARAMETER VALUES FOR EXPECTED CASUALTY RATES

ρ	μ_0	μ_1
.01	1 95	2.33
.02	.94	2 26
.05	.35	2.06
.10	.16	1.80
25	051	1.30

$$\begin{aligned}\bar{B} &\equiv E(B_{t+1} \mid B_{t+1} > 0, B_t = 0) \\ &= \frac{1}{5}(.005 + .23 + .004 + .071 + .001) = .062.\end{aligned} \tag{A9}$$

Since war could break out at any time during the year, the annualized value of EB_{t+1}, denoted by $\bar{B}^A$, would be roughly twice the above figure—that is, $\bar{B}^A \approx .124$.

Finally, when B_{t+1} and B_t are both positive, the conditional expectation for B_t is given by

$$E(B_{t+1} \mid B_{t+1} > 0, B_t > 0) = \theta_0 + \theta_1 B_t^A,$$

where B_t^A is the current casualty rate expressed at an annual rate if hostilities applied only to a fraction of year t. The parameter θ_1 is based on the assumption (not refuted by the small sample of U.S. data) that wars tend neither to grow nor to contract over time, except that war may end at some time during year $t + 1$ as governed by the parameter p_2. Accordingly, $\theta_1 \approx 1 - \frac{1}{2}(1 - p_2) = .90$. The parameter θ_0 is set so that the value of the Ψ_t variable, which appears in equation (A6), converges to the value associated with $B_{t+2} = 0$ as $B_{t+2} \to 0$ (which essentially recognizes that a new war may break out next year, even if one is already going on). The value of θ_0 turns out to be $\theta_0 \approx p_1 \bar{B}^A = .007$. Accordingly, I use the relation

$$E(B_{t+1} \mid B_{t+1} > 0, B_t > 0) = .007 + .90 B_t^A. \tag{A10}$$

Equations (A7)–(A10) allow calculation of the relevant expectation of future casualty rates, which appears in equation (A6), $\Psi_t \equiv \Sigma_{i=1}^{\infty} EB_{t+i+2}/(1 + \rho)^i$, conditional on observation of B through B_{t+2} and for a given value of the discount rate ρ. The result takes the form

$$\Psi_t = \mu_0 + \mu_1 B_{t+2}^A, \tag{A11}$$

where μ_0 and μ_1 can be determined as functions of the ρ parameter.[37] Specifically, these coefficients for selected values of ρ can be seen in table A1. Since ρ corresponds to the difference between the real rate of return and the real growth rate, the values of the μ coefficients corresponding to the lower

[37] The general formulae are:

$$\begin{aligned}\mu_0 &= [\bar{B}p_1(1 - p_2)(1 + \rho + p_2\theta_1) \\ &\quad + p_2\theta_0(1 + \rho)(\rho + p_1)]/[\rho(1 + \rho + p_1 - p_2)(1 + \rho - p_2\theta_1)], \\ \mu_1 &= p_2\theta_1/(1 + \rho - p_2\theta_1).\end{aligned}$$

values of ρ would seem to be most pertinent. The empirical results that are reported in the text use the value $\rho = .02$ per year.

The combination of equation (A11) with equations (A5) and (A6) allows calculation of the perceived permanent government purchases ratio, $\overline{g_t^p}$, as a function of the variables $(g_t^p, k_{t-1}^p, B_{t-3}, \ldots, B_{t+2})$ and the parameters $(\rho, \gamma, \alpha_0, \alpha_1, \alpha_2, \alpha_3, a_1, a_2)$, where ρ is the net real discount rate, γ measures the reaction of current defense purchases to existing capital stock, and the α's and a's describe the effect of the array of B variables on defense purchases. The results are therefore expressed in terms of the general form of equation (6) in the text. Other coefficients that appear in the analysis (δ', b, p_1, p_2, $\bar{B}$—see, e.g., the expressions contained in n. 37 above) are treated as fixed at the values: $\delta' = .16$ per year, $b = .34$, $p_1 = .06$, $p_2 = .79$, and $\bar{B} = .062$.

References

Bailey, Martin J. *National Income and the Price Level: A Study in Macroeconomic Theory.* 2d ed. New York: McGraw-Hill, 1971.

Barro, Robert J. "Public Debt and Taxes." In *Federal Tax Reform: Myths and Realities,* edited by Michael Boskin. San Francisco: Inst. Contemporary Studies, 1978.

———. "On the Determination of the Public Debt." *J.P.E.* 87, no. 5, pt. 1 (October 1979): 940–71.

———. "A Capital Market in an Equilibrium Business Cycle Model." *Econometrica* 48 (September 1980): 1393–1417. (*a*)

———. "Federal Deficit Policy and the Effects of Public Debt Shocks." *J. Money, Credit and Banking* 12, no. 4, pt. 2 (November 1980): 747–62. (*b*)

———. "Intertemporal Substitution and the Business Cycle." *Carnegie-Rochester Conference Series on Public Policy* 14 (Spring 1981): 237–68. (*a*)

———. "Unanticipated Money Growth and Economic Activity in the United States." In *Money, Expectations, and Business Cycles: Essays in Macroeconomics,* by Robert J. Barro. New York: Academic Press, 1981. (*b*)

Benjamin, Daniel K., and Kochin, Levis A. "A Theory of State and Local Government Debt." Unpublished paper, Univ. Washington, 1978.

Hall, Robert E. "Labor Supply and Aggregate Fluctuations." *Carnegie-Rochester Conference Series on Public Policy* 12 (Spring 1980): 7–33.

Kendrick, John W. *Productivity Trends in the United States.* Princeton, N.J.: Princeton Univ. Press (for Nat. Bur. Econ. Res.), 1961.

———. *The Formation and Stocks of Total Capital.* New York: Columbia Univ. Press (for Nat. Bur. Econ. Res.), 1976.

Kydland, Finn E., and Prescott, Edward C. "Rules Rather than Discretion: The Inconsistency of Optimal Plans." *J.P.E.* 85, no. 3 (June 1977): 473–91.

———. "A Competitive Theory of Fluctuations and the Feasibility and Desirability of Stabilization Policy." In *Rational Expectations and Economic Policy,* edited by Stanley Fischer. Chicago: Univ. Chicago Press (for Nat. Bur. Econ. Res.), 1980.

Kuznets, Simon. "Discussion of the New Department of Commerce Income Series: National Income: A New Version." *Rev. Econ. and Stats.* 30 (August 1948): 151–79.

Lucas, Robert E., Jr. "An Equilibrium Model of the Business Cycle." *J.P.E.* 83, no. 6 (December 1975): 1113–44.

Lucas, Robert E., Jr., and Rapping, Leonard A. "Real Wages, Employment, and Inflation." *J.P.E.* 77, no. 5 (September/October 1969): 721–54.

Musgrave, Richard A. *The Theory of Public Finance: Study in Public Economy.* New York: McGraw-Hill, 1959.

Plosser, Charles I., and Schwert, G. William. "Money, Income, and Sunspots: Measuring Economic Relationships and the Effects of Differencing." *J. Monetary Econ.* 4 (November 1978): 637–60.

Samuelson, Paul A. "The Pure Theory of Public Expenditure." *Rev. Econ. and Statis.* 36 (November 1954): 387–89.

Thompson, Earl A. "Taxation and National Defense." *J.P.E.* 82, no. 4 (July/August 1974): 755–82.

Information Costs, Duration of Search, and Turnover: Theory and Applications

Louis L. Wilde

California Institute of Technology

This paper uses a formal model of search over multiattribute alternatives, analyzed in a product market setting, to investigate the theoretical foundations of the empirical literature on duration of search and turnover in product markets, labor markets, and marriage markets. A number of specific empirical predictions are also derived. In particular, whether "quality" is a "search" attribute or an "experience" attribute is related to the cost of search, the cost of inspection, the price of the good, and certain properties of the market distribution of price and quality

I. Introduction

Several years ago Phillip Nelson observed that certain problems arise in extending search theory to deal with nonhomogeneous goods. In particular, he noted that "information about quality differs from information about price because the former is usually more expensive to buy than the latter" (1970, p. 311). To analyze the implications of this observation, Nelson divided goods into two classes, search goods and experience goods. Search goods are those for which utility is assessed before purchase by actual inspection. Experience goods are those for which utility is assessed after purchase by actual consumption.

This paper is a substantially revised version of "Further Results on Inspection and Evaluation in Product Markets" (unpublished manuscript, California Institute of Technology, February 1978). I would like to thank Steve Lippman and Alan Schwartz for very helpful comments on an earlier draft. Comparative statics results are proved in an Appendix which is available from the author (California Institute of Technology, Social Science Working Paper no. 306, January 1980).

[*Journal of Political Economy*, 1981, vol. 89, no. 6]
© 1981 by The University of Chicago 0022-3808/81/8906-0012$01.50

These definitions turn out to be very strong. So strong, in fact, that they lead to some difficulties. Consider, for example, experience goods. For these goods utility is assessed after purchase by actual consumption. But at least the price is observed before purchase. Since for experience goods this cannot, by definition, affect the choice of which brand to buy, Nelson was forced to assume that "consumers either sample at random from among all brands or from among those brands in the price range the consumer deems appropriate for himself" (1970, p. 313). These assumptions require that consumers either ignore prices completely or have perfect information regarding prices, neither of which seems likely.

The problem with Nelson's definitions is that goods generally possess a number of characteristics which can differ in their degree of observability. Thus a good might possess some "search characteristics" and some "experience characteristics." Furthermore, whether a particular characteristic is a search characteristic or an experience characteristic ought to be endogenously determined by the consumer.

In previous work, I have analyzed a model in which goods are described by precisely two characteristics, price and quality. The market is described by an exogenous distribution of these characteristics. Sampling from this distribution is assumed to be costly. Once an observation is drawn, price is observed costlessly. Quality, however, can never be observed before purchase. In this case quality is an extreme example of an experience characteristic, one for which the cost of observation prior to purchase is infinite. The present paper extends this model to allow quality to be observed at some finite cost. The purpose is to explicate the relationships between information costs, duration of search, and turnover.[1]

This model, introduced in Wilde (1980), applies to a number of cases discussed by Nelson. For example, it formalizes his prototypic experience good, canned tuna fish. Nelson suggested that "to evaluate brands of canned tuna fish, for example, the consumer would almost certainly purchase brands of tuna fish for consumption. He could, then, determine from several purchases which brand he preferred. For tuna fish there is no effective search alternative open. At the low price of experience, there is insufficient demand for specialized establishments selling tastes of various brands of tuna fish" (1970, p. 312). In this case, Nelson seems to be suggesting that the "price of experience" is low because the price of the good is low. However, the

[1] The formal analysis in this paper will be set in the product market. In this case turnover is a nonrepeat sale. Sec. IV will discuss applications to the labor market and the marriage market. In the labor market, turnover is a quit and in the marriage market turnover is a divorce.

price of experience can be low for other reasons as well. For example, if there is little variance in quality, then the price of experience is low because there is little chance of purchasing a low-quality good. In fact, it will turn out that by allowing quality to be observed before purchase an explicit expression for the price of experience can be derived. Utilizing this expression, this paper will also explicate the relationship between information costs and the price of experience.

The focus of the model developed in Section II is on an imperfectly informed consumer who is interested in maintaining a flow of consumption of a good which is described by price and quality. The market offers various combinations of price and quality, but the consumer cannot costlessly observe them; by paying a search cost he or she can sample a good from the market, but only price is observed. Quality can be observed either before purchase by actual inspection (at some additional cost) or after purchase by actual consumption. Whether quality is observed before purchase (herein called inspection) or after purchase (herein called evaluation), the consumer can return to the market and resample if the observed quality is too low. The initial problem is to characterize the optimal strategy for a consumer in such an environment. For a fixed utility function, joint distribution of price and quality, and cost of search, three possibilities arise depending on the cost of inspection: (1) If the cost of inspection is low enough, inspection will be the optimal strategy for low prices and drawing a new observation will be the optimal strategy for high prices. (2) If the cost of inspection is of an intermediate amount, evaluation will be the optimal strategy for low prices, inspection will be the optimal strategy for intermediate prices, and drawing a new observation will be the optimal strategy for high prices. (3) If the cost of inspection is high enough, evaluation will be the optimal strategy for low prices and drawing a new observation will be the optimal strategy for high prices.

The formal results of the paper relate to the characterization of the consumer's optimal strategy which is summarized by statements 1–3 above. However, these results have broad implications. Three important general observations emerge from the analysis.

First, as mentioned above, the price of experience can be defined analytically. This is important because it allows one to differentiate between the direct, short-run benefits of purchasing a good which are derived from its consumption and the indirect, long-run benefits of purchasing a good which are derived from evaluation of its quality attributes.

Second, it will be demonstrated that in some cases quality will be a pure search characteristic (statement 1 above), in some cases it will be

neither a pure search characteristic nor a pure experience characteristic (statement 2 above), and in some cases it will be a pure experience characteristic (statement 3 above). Hence, not only is determining whether a particular good is an experience good or a search good a complex matter, but so is determining whether a particular characteristic of that good is an experience characteristic or a search characteristic. Since these distinctions have become very popular in the literature it is important to understand their limitations.

Finally, the comparative statics associated with the characterization of the consumer's optimal strategy will show that the theoretical foundations of much of the empirical work on duration of search and turnover in labor markets and marriage markets are unsound.

A number of specific empirical predictions are also derived. It is shown that quality is more likely to be a search attribute (1) the lower is the cost of search, (2) the lower is the cost of inspection, (3) the higher is the price of the good, and (4) the larger is the variation in utility due to quality relative to the variation in utility due to price. Similarly, it is shown that quality is more likely to be an experience attribute (1) the higher is the cost of search, (2) the higher is the cost of inspection, (3) the lower is the price of the good, and (4) the smaller is the variation in utility due to quality relative to the variation in utility due to price.

It is also shown that regardless of whether quality is a search attribute or an experience attribute, increases in search costs decrease the duration of search. In general, increases in inspection costs have ambiguous effects on the duration of search, but it is argued that if the variation in utility due to quality relative to the variation in utility due to price is low, increases in inspection costs are likely to increase the duration of search (at least by one measure).

Finally, considerable attention is focused on turnover. Again, in the most general setting results tend to be ambiguous. This is especially true regarding the relationship between search costs and turnover. However, it is argued that if the variation in utility due to quality relative to the variation in utility due to price is low, increases in inspection costs are likely to increase turnover.

This paper is organized as follows. Section II introduces the basic model. Section III considers the three cases mentioned above: (*a*) in which quality is always observed before purchase; (*b*) in which quality is observed before purchase for some prices but is observed after purchase for other prices; and (*c*) in which quality is always observed after purchase. Section IV discusses the empirical implications of the model in the labor market and the marriage market. Section V provides a brief conclusion.

II. The Model: Notation, Definitions, and Assumptions

In this section the basic model will be developed and an analytical expression for the price of experience derived. Section III provides comparative statics for the three cases mentioned in the introduction.

Assume the good which is sought by the consumer has a lifetime of one period. Let $U(p, q)$ be the total net value to the consumer of purchasing and consuming the good characterized by price p and quality q, where U is differentiable and bounded on $R_+ \times R_+$ with $\partial U/\partial p < 0$ and $\partial U/\partial q > 0$. Let $\phi(p, q)$ be the consumer's subjective estimate of the market density of P and Q. For mathematical convenience, assume ϕ is strictly positive on $R_+ \times R_+$. Define $f(p)$ as the marginal density of P and $g(q \mid p)$ as the conditional density of Q given $P = p$, both based on ϕ.

The cost of drawing an observation at random from ϕ is c_S, where $c_S \geqslant 0$. The cost of observing the true value of Q prior to purchase is c_T, where $c_T \geqslant 0$ (both c_S and c_T are measured in the same units as U).

The consumer can sample as many observations as desired from ϕ at the beginning of each period. Any number of inspections are also allowed. However, the consumer demands precisely one unit of the good each period.[2] The consumer's objective is to maximize his or her expected discounted utility of consumption net of search costs. Sampling is assumed to be without recall, the horizon is infinite, and the discount rate is β, where $0 < \beta < 1$.

Now suppose the consumer has sampled a good priced at p. Three reactions are possible: sample again without observing quality; inspect quality and then either buy the good forever or sample again; or evaluate quality and then either buy the good forever or sample again. Let $v(p)$ be the expected value of drawing an observation of p and then proceeding optimally. Then

$$v(p) = -c_S + \max\{V, B(p), T(p)\}, \tag{1}$$

where V is the expected value of search (i.e., the expected value of $v[p]$ taken with respect to f), $B(p)$ is the expected value of buying the good without observing quality and then proceeding optimally, and $T(p)$ is the expected value of testing quality prior to purchasing the good and then proceeding optimally.

To define $B(p)$ and $T(p)$ analytically, observe first that, once quality is known, the value of the optimal policy is the same whether quality is observed via inspection or evaluation. In other words, once quality has been observed, it is irrelevant how it was observed—any asso-

[2] Search and inspection are assumed to be timeless in order to avoid confounding the direct costs of these activities with the opportunity cost of delaying the purchase decision.

ciated costs will by then be sunk costs. Define the value of the optimal policy once quality is known as $k(p)$. Then

$$k(p) = VG[q^*(p) \mid p] + \int_{q^*(p)}^{\infty} \frac{U(p, q)}{1 - \beta} g(q \mid p) dq, \tag{2}$$

where $q^*(p)$ is that quality level which makes the consumer indifferent between consuming the good characterized by $[p, q^*(p)]$ and searching again for a new good from ϕ. That is, $q^*(p)$ is defined by $U[p, q^*(p)] = V(1 - \beta)$. The logic of (2) is that if $q < q^*(p)$, then the consumer rejects the good and samples again, receiving V (the expected value of search). This happens with probability $G[q^*(p) \mid p]$. If $q \geq q^*(p)$, then the good is acceptable and the consumer receives the conditional expected value of $U(p, q)/(1 - \beta)$, given $q \geq q^*(p)$.

Using $k(p)$, both $B(p)$ and $T(p)$ are defined straightforwardly. If the consumer buys the good priced at p, he or she receives the expected utility of consumption, given quality is unknown, plus the discounted value of an optimal policy once quality is known. Hence

$$B(p) = EU(p, Q) + \beta k(p). \tag{3}$$

If the consumer tests for quality prior to purchase, he or she pays c_T and receives the undiscounted value of an optimal policy once quality is known. Hence

$$T(p) = -c_T + k(p). \tag{4}$$

From (3), by adding and subtracting $k(p)$ and rearranging, $B(p)$ can be rewritten as $B(p) = -[(1 - \beta)k(p) - EU(p, Q)] + k(p)$. The term in brackets is a cost which is directly comparable to c_T. That is, one can define

$$c_B(p) = (1 - \beta)k(p) - EU(p, Q). \tag{5}$$

Then $B(p) = -c_B(p) + k(p)$.

Equation (5) provides an analytical expression for Nelson's price of experience. Furthermore, it has a natural interpretation. Recall that $k(p)$ is the expected value of an optimal policy once quality is known, given the observed price is p. When quality is observed via actual consumption, this value is not obtained for one period (since in this model the good lasts for precisely one period). Hence $(1 - \beta)k(p)$ is the gross opportunity cost of consuming the good, given quality is unknown. But consumption of the good yields utility, in this case $EU(p, Q)$. The net opportunity cost of consuming the good, given quality is unknown, is the difference between these two quantities.

In analyzing the optimal policy it will be convenient to make a transformation of variables in the definition of $k(p)$. Since it is ulti-

mately final utility which matters to the consumer, the focus of (2) can be shifted from the conditional distribution of quality given price to the conditional distribution of utility given price. That is, let $\Psi(w \mid p)$ be the conditional distribution of utility given $P = p$. Then $k(p)$ becomes

$$k(p) = V\Psi[V(1-\beta) \mid p] + \int_{V(1-\beta)}^{\bar{z}(p)} \frac{w}{1-\beta} \psi(w \mid p) dw, \tag{6}$$

where $\bar{z}(p) = \lim_{q \to \infty} U(p, q)$. Integrating (6) by parts gives

$$k(p) = (1-\beta)^{-1}\left[\bar{z}(p) - \int_{V(1-\beta)}^{\bar{z}(p)} \Psi(w \mid p) dw\right]. \tag{7}$$

This form of $k(p)$ is intuitively less appealing than (6), but it is more powerful. For example, using (5) it is easy to show from (7) that

$$c_B(p) = \int_{\bar{z}(p)}^{V(1-\beta)} \Psi(w \mid p) dw. \tag{8}$$

where $\bar{z}(p) = U(p, 0)$. Expressing the price of experience in this form is useful analytically since it makes the comparison between $B(p)$ and $T(p)$ easier. Conceptually, it helps identify factors which might affect the decision whether to observe quality before purchase or after purchase. For example, suppose the cost of search, c_S, increases. Then surely the value of an optimal policy will fall, that is, $\partial V/\partial c_S < 0$. Equation (8) suggests that the price of experience will then fall as well. Hence the higher is the cost of search, the more likely it is that evaluation will be optimal. These and other results will be formalized in the following section. First, however, a few more preliminary assumptions will be needed.

Using the definition of $c_B(p)$ introduced above in equation (5), the functional equation (1) can be rewritten as

$$v(p) = -c_S + \max\{V, k(p) - c_B(p), k(p) - c_T\}.^3 \tag{9}$$

The next step in characterizing an optimal policy is to compare V, $k(p) - c_B(p)$, and $k(p) - c_T$. Unfortunately, without more structure on the joint distribution of price and quality, any number of things can happen. Nelson recognized this problem as well, stating the following: "Prior to using [a] brand, all the consumer knows is its price. But this knowledge provides only the roughest sort of guide to choice, for the consumer must assume a generally positive relationship between price

[3] It is convenient at this point to assume that the functional equation (1) has a unique, bounded solution. In this case V is unique and well defined. Existence of a bounded solution to (1) is straightforward. Uniqueness can be established along traditional lines if search and inspection are not assumed to be timeless (see Wilde 1979) or by appealing to the appropriate generalization of MacQueen and Miller (1960).

and quality. In the absence of any other information, the consumer would not know if he were better off experimenting with low- or high-priced brands" (1970, p. 313). To get around this problem, Nelson converts the joint distribution of price and quality to a distribution of net utility and proceeds under the assumption that evaluation is always used to observe quality. In the present analysis, since the decision whether to observe quality before purchase or after purchase is endogenous, a more formal approach is needed. The standard assumption is that $\partial\Psi(w \mid p)/\partial p > 0$. This assumption implies that, on average, higher price is associated with lower utility even though higher price may well be associated with higher quality. Because this assumption is discussed at length in Wilde (1980), it will be assumed to hold here without further rationalization.[4]

Several implications follow directly from the assumption that $\partial\Psi(w \mid p)/\partial p > 0$. First, it implies that the price of experience is increasing in the price of the good. To see this, simply take the derivative of (8) with respect to p:

$$\begin{aligned} c_B'(p) &= \int_{\bar{z}(p)}^{V(1-\beta)} [\partial\Psi(w \mid p)/\partial p]dw - \Psi[\bar{z}(p) \mid p]\bar{z}'(p) \\ &= \int_{\bar{z}(p)}^{V(1-\beta)} [\partial\Psi(w \mid p)/\partial p]dw, \end{aligned}$$

since $\Psi[\bar{z}(p) \mid p] = 0$ by definition. Second, $\partial\Psi(w \mid p)/\partial p > 0$ implies that observing quality before purchase becomes a less desirable alternative to sampling again as the observed price increases. That is, using (6),

$$T'(p) = k'(p) = -\int_{V(1-\beta)}^{\bar{z}(p)} [\partial\Psi(w \mid p)/\partial p]dw \leq 0. \tag{10}$$

Moreover, $B'(p) = k'(p) - c_B'(p) \leq T'(p)$ since $c_B'(p) \geq 0$. Hence observing quality after purchase also becomes a less desirable alternative to sampling again as the observed price increases, and it does so at an even faster rate than observing quality before purchase.

III. The Optimal Policy

It turns out that three qualitatively distinct forms of the optimal policy are possible. If c_T is low enough, then inspection always dominates evaluation. If c_T is somewhat higher, then evaluation dominates in-

[4] Even in the case where c_T is infinite, some formal structure must be placed on Ψ. See Wilde (1980) or Hey and McKenna (1981) for more details.

spection for one set of prices and inspection dominates evaluation for another set of prices; and, if c_T is high enough, then evaluation always dominates inspection. These three cases are analyzed next.

A. *Inspection Only*

In comparing the expected value of observing quality before purchase to the expected value of observing quality after purchase, the crucial parameters are c_T and $c_B(p)$, the respective costs of these two activities. It is clear from (3) and (4) that $B(p) \gtreqless T(p)$ as $c_T \gtreqless c_B(p)$. But $c_B(p)$ is increasing in p so that $c_B(0) \geq c_T$ implies $c_B(p) \geq c_T$ for all $p \geq 0$. Hence the expected value of observing quality before purchase will always be greater than the expected value of observing quality after purchase when $c_B(0) > c_T$. This is obviously most likely to be the case when $c_B(0)$ is large and c_T is small. Equation (8) suggests that $c_B(0)$ is most likely to be large when $V(1 - \beta)$ is significantly greater than $\bar{z}(0)$. But $V(1 - \beta) = U[0, q^*(0)]$ and $\bar{z}(0) = U(0, 0)$. Hence $c_B(0)$ is most likely to be large when $q^*(0)$ is high. In other words, inspection is likely to dominate evaluation for all prices when the cost of inspection is low or when few quality levels are acceptable even at low prices. The latter is the case, for example, when the cost of drawing observations from ϕ is low or the variation in utility due to quality is high relative to the variation in utility due to price.[5]

In the remainder of this subsection it will be assumed that $c_B(0) \geq c_T$; that is, it will be assumed that

$$\int_{\bar{z}(0)}^{V(1-\beta)} \Psi(w \mid 0)dw \geq c_T.$$

This implies that $B(p) \leq T(p)$ for all $p \geq 0$, in which case $B(p)$ can be ignored completely; characterizing the optimal policy reduces to comparing $T(p)$ and V. Two possibilities arise. In the first there exists a unique finite price, say p_T^*, such that observing quality prior to purchase is optimal for $p \leq p_T^*$ and sampling again is optimal for $p \geq p_T^*$ (see fig. 1). The critical price is defined by $V = T(p_T^*)$.[6] In the second, V is strictly less than $T(p)$ for all $p \geq 0$ so that observing quality prior to purchase is always optimal.[7]

[5] There is no guarantee that $c_B(p) > 0$. In particular, it might be the case that $c_B(0) = 0$. It can be shown, however, that there exists $\epsilon > 0$ such that $c_B(0) > 0$ if $c_S < \epsilon$.

[6] Since $\partial\Psi(w \mid p)/\partial p > 0$ only implies $T'(p) \leq 0$, it is possible that the equation $T(p) = V$ does not have a unique solution. However, if this is the case, then $\{p \mid T(p) = V\} = [p_T^*, \infty)$, where $p_T^* = \inf\{p \mid T(p) = V\}$. If p_T^* is defined in this fashion, it satisfies the formal requirements stated in the text.

[7] Formal proofs of these assertions have been omitted since they are trivial. It is apparent, however, that since $c_B(0) \geq 0$, there is always a small enough value of c_T (possibly 0) such that $B(p) \leq T(p)$ for all $p \geq 0$.

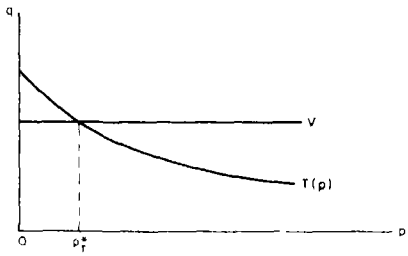


FIG. 1.—Definition of p_T^*

Assume that p_T^* exists and is finite. Then p_T^* and q^* partition $R_+ \times R_+$ into three sets (see fig. 2). In region I, $p > p_T^*$, so the good is rejected outright. In region II, $p \leqslant p_T^*$ but $q < q^*(p)$, so that quality is observed prior to purchase but the good is subsequently rejected (i.e., not purchased). In region III, $p \leqslant p_T^*$ and $q \geqslant q^*(p)$, so that quality is observed prior to purchase and the good is subsequently accepted (i.e., purchased).

How do changes in c_S and c_T affect this partition? To answer this question one needs to know how changes in c_S and c_T affect p_T^* and q^*. The following results are straightforward but tedious and can be found in the (unpublished) Appendix to this paper. It is shown there that $dp_T^*/dc_S > 0$ and $dp_T^*/dc_T < 0$. Of course V falls as either c_S or c_T rises. Since $U[p, q^*(p)] = V(1 - \beta)$ by definition and U is increasing in q, this implies $dq^*(p)/dc_S < 0$ and $dq^*(p)/dc_T < 0$. In other words, an increase in the cost of drawing observations from ϕ will make inspection an optimal strategy for more prices while an increase in the cost of inspection will make inspection an optimal strategy for less prices. An increase in either cost will make more quality levels acceptable for any given price.

Next, consider how changes in c_S and c_T affect the number of observations which must be drawn from ϕ (whether or not quality is inspected) before an acceptable good is found. Define

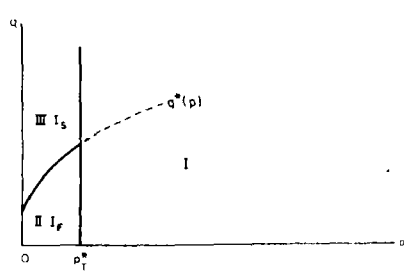


FIG. 2.—Price-quality combinations, case A: I = reject outright, II = inspect and subsequently reject, III = inspect and subsequently accept.

$$I_S = \int_0^{p_T^*} \{1 - G[q^*(p) \mid p]\} f(p) dp$$

and

$$I_F = \int_0^{p_T^*} G[q^*(p) \mid p] f(p) dp.$$

Here I_F is the probability that a random price-quality combination will be inspected and rejected, and I_S is the probability that a random price-quality combination will be inspected and accepted. It is again shown in the Appendix that the following hold: $\partial I_S/\partial c_S > 0$ and $\partial I_F/\partial c_S \gtreqless 0$, while $\partial I_S/\partial c_T \gtreqless 0$ and $\partial I_F/\partial c_T < 0$.

Consider first an increase in the cost of drawing observations. Since p_T^* increases, inspection becomes an optimal strategy for more prices. Furthermore, since $q^*(p)$ decreases, more quality levels are acceptable for any given price. Hence the expected number of observations which must be drawn from ϕ before an acceptable good is found $(1/I_S)$ falls. The effect of an increase in c_S on I_F is ambiguous, however, because the decrease in $q^*(p)$ counteracts the increase in p_T^* rather than reinforcing it.

Precisely the opposite happens when the cost of inspection increases. Since p_T^* falls, inspection becomes an optimal strategy for less prices. Furthermore, since $q^*(p)$ still decreases, fewer quality levels are acceptable for any given price. Hence I_F falls. Since the effect on I_S is ambiguous, it is impossible to assert that an increase in the cost of inspection reduces the expected number of observations which must be drawn from ϕ before an acceptable good is found.

B. *Inspection and Evaluation*

It was assumed throughout Section IIIA that $c_B(0) \geqslant c_T$ so that inspection dominated evaluation for all prices. Recall that

$$c_B(0) = \int_{\bar{z}(0)}^{V(1-\beta)} \Psi(w \mid 0) dw.$$

It is apparent from this equation that $\partial c_B(0)/\partial c_S < 0$ and $\partial c_B(0)/\partial c_T < 0$ since V is decreasing in either cost and $\Psi(w \mid 0) > 0$ for w close to $\bar{z}(0)$. Hence increases in either c_S or c_T make it less likely that inspection will dominate evaluation for all prices.

Assume that $c_B(0) < c_T$. Then two possibilities arise. In the first, inspection is optimal for one set of prices and evaluation is optimal for another set of prices. In the second, inspection is never optimal. The first possibility will be analyzed now and the second will be considered in Section IIIC.[8]

[8] Again, formal proofs that ranges of c_S and c_T exist such that both possibilities occur have been omitted. The text following identifies two necessary conditions for inspection

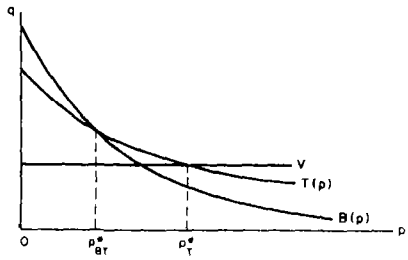


FIG. 3.—Definition of p^*_{BT} and p^*_T

In order for inspection to be optimal for one set of prices and evaluation to be optimal for another set of prices two conditions must be met. Assuming $c_B(0) < c_T$, the first can be stated as follows.

> CONDITION 1: The cost of inspection must not be so great that evaluation dominates inspection for all prices.

When condition 1 holds, since $B'(p) \leq T'(p) \leq 0$, there exists a finite price p^*_{BT} such that evaluation dominates inspection for $p \leq p^*_{BT}$ and inspection dominates evaluation for $p \geq p^*_{BT}$. The critical price is defined by $T(p^*_{BT}) = B(p^*_{BT})$ (see fig. 3).

Assuming condition 1 holds, the second condition can be stated as follows.

> CONDITION 2: The expected return to search must be low enough that for some prices inspection dominates drawing another observation from ϕ.

Condition 2 requires that for some $p > p^*_{BT}$, $T(p) > V$. However, as in case A, it might be that $T(p) > V$ for all $p \geq p^*_{BT}$. Assume this does not hold. Then there exists a finite price p^*_T such that inspection is optimal for $p \in [p^*_{BT}, p^*_T]$, and drawing another observation from ϕ is optimal for $p \geq p^*_T$. Again, as in Section IIIA, the critical price is defined by $T(p^*_T) = V$ (see fig. 3).[9]

Overall, the situation dealt with in this case is the most interesting of the model because it shows that quality can be a search characteristic for some prices and an experience characteristic for other prices. That is, the configuration of utility, search costs, inspection costs, and the joint distribution of price and quality are such that inspection, evaluation, and drawing another observation from ϕ are all optimal strategies for various prices. In general, p^*_T, p^*_{BT}, and q^* partition $R_+ \times R_+$ into five regions (see fig. 4). As in case A, region I includes prices

to be optimal for one set of prices and evaluation to be optimal for another set of prices. There seems to be little value in making these more formal.

[9] Again, since $\partial\Psi(w|p)/\partial p > 0$ only implies $B'(p) \leq 0$, the solution to the equation $T(p) = B(p)$ may not be unique. In this case $\{p|T(p) = B(p)\} \equiv [p^*_{BT}, \infty)$ where $p^*_{BT} = \inf \{p|T(p) = B(p)\}$. If p^*_{BT} is defined in this fashion, it satisfies the formal requirements stated in the text.

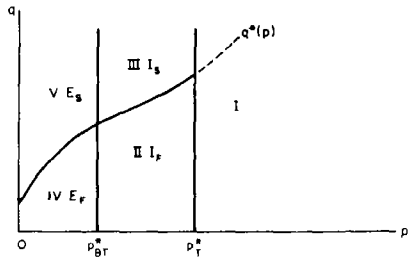


FIG. 4.—Price-quality combinations, case B: I = reject outright, II = inspect and ultimately reject, III = inspect and ultimately accept, IV = purchase once and only once, V = purchase once and forevermore.

for which rejecting the good outright is optimal. In region II, $p^*_{BT} < p \leq p^*_T$ but $q < q^*(p)$, so that quality is observed prior to purchase but the good is subsequently rejected; and in region III, $p^*_{BT} < p \leq p^*_T$, but $q \geq q^*(p)$, so that quality is observed prior to purchase and the good is subsequently accepted. There are two additional regions, though. In region IV, $p \leq p^*_{BT}$ and $q < q^*(p)$, so that the good is purchased without quality having been observed but is not repurchased. In region V, $p \leq p^*_{BT}$ and $q \geq q^*(p)$, so that the good is purchased without quality having been observed and is repurchased in all subsequent periods.

How do changes in c_S and c_T affect this partition? As before, to answer this question we need to know how changes in c_S and c_T affect p^*_{BT}, p^*_T, and $q^*(p)$. As in case A, $dp^*_T/dc_S > 0$ and $dp^*_T/dc_T < 0$. Also, $dp^*_{BT}/dc_S > 0$ and $dp^*_{BT}/dc_T > 0$. Finally, $dq^*(p)/dc_S < 0$ and $dq^*(p)/dc_T < 0$.

An increase in the cost of drawing observations will increase the set of prices for which evaluation is an optimal strategy. Furthermore, it will increase the set of prices for which either inspection or evaluation is an optimal strategy. An increase in the cost of inspection will increase the set of prices for which evaluation is an optimal strategy, will decrease the set of prices for which inspection is an optimal strategy, and will decrease the set of prices for which either inspection or evaluation is an optimal strategy. Finally, an increase in either c_S or c_T will make more quality levels acceptable at any given price.

It is again useful to consider how increases in c_S and c_T affect the probability that a random price-quantity combination will fall in any given region. Define

$$E_S = \int_0^{p^*_{BT}} \{1 - G[q^*(p) \mid p]\} f(p)\, dp,$$

$$E_F = \int_0^{p^*_{BT}} G[q^*(p) \mid p] f(p) dp,$$

$$I_S = \int_{p_{BT}^*}^{p_T^*} \{1 - G[q^*(p) \mid p]\} f(p) dp,$$

and

$$I_F = \int_{p_{BT}^*}^{p_T^*} G[q^*(p) \mid p] f(p) dp.$$

Here I_F, I_S, E_F, and E_S are the probabilities a random price-quality combination will fall in region II, III, IV, or V, respectively. It can be shown that

$$\frac{\partial I_S}{\partial c_S} \gtreqless 0, \frac{\partial I_F}{\partial c_S} \gtreqless 0, \frac{\partial E_S}{\partial c_S} > 0, \quad \text{and} \quad \frac{\partial E_F}{\partial c_S} \gtreqless 0.$$

Furthermore,

$$\frac{\partial I_S}{\partial c_T} \gtreqless 0, \frac{\partial I_F}{\partial c_T} < 0, \frac{\partial E_S}{\partial c_T} > 0, \quad \text{and} \quad \frac{\partial E_F}{\partial c_T} \gtreqless 0.$$

Finally,

$$\frac{\partial (I_S + E_S)}{\partial c_S} > 0 \quad \text{and} \quad \frac{\partial (I_S + E_S)}{\partial c_T} \gtreqless 0.$$

While the majority of these partial derivatives are ambiguous in sign, a number of interesting observations can still be made. First, the probability that a random price-quality combination will be acceptable (whether quality is inspected or evaluated) is given by $I_S + E_S$. Hence the expected number of observations needed to locate an acceptable good is $1/(I_S + E_S)$. As before, this quantity is decreasing in c_S and ambiguous in c_T. Second, both $\partial E_S/\partial c_S > 0$ and $\partial E_S/\partial c_T > 0$. That is, the probability that a random price-quality combination will be purchased without quality having been observed and subsequently repurchased is increasing in either cost. This is because when either c_S or c_T increases, the set of prices for which evaluation is optimal increases and the set of quality levels which are acceptable for any given price also increases. However, this necessarily implies $\partial E_F/\partial c_S$ and $\partial E_F/\partial c_T$ are ambiguous in sign. This last observation is important. When a price-quality combination falls in region IV, the good is purchased without quality having been observed but is not subsequently repurchased. This "brand disloyalty" is analogous to a job quit in the labor market or a divorce in the marriage market. It is crucial to recognize that the likelihood of these events does not appear to be systematically related to either search costs or inspection costs. This point will be discussed in more detail in Section IV.

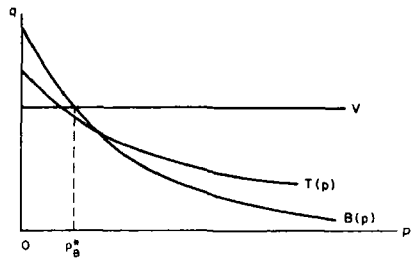


FIG. 5.—Definition of p_B^*

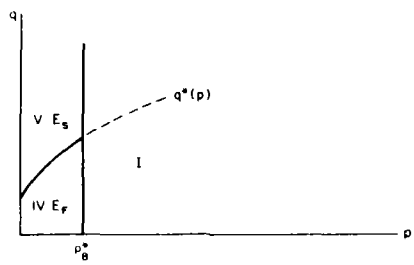


FIG. 6.—Price-quality combinations, case C: I = reject outright, IV = purchase once and only once, V = purchase once and forevermore.

C. *Evaluation Only*

The final case is one in which $c_B(0) < c_T$, but one of the conditions of case B does not hold.[10] Under these circumstances evaluation will always dominate inspection for any acceptable price.[11] Both this case and case A are, in a sense, special cases of case B. Here the relevant critical price, p_B^*, is defined by $B(p_B^*) = V$ (see fig. 5). The elements of the associated partition correspond to regions I, IV, and V of case B (see fig. 6). Of course c_T has no effect on this partition. The effects of c_S are identical to those in case B (where p_B^* now replaces p_{BT}^*).

IV. Applications

The model analyzed in this paper has obvious analogues in the labor market and the marriage market. The product market has been used as the setting to this point because much of the relevant literature deals with consumer behavior (e.g., Nelson 1970; Lippman and

[10] This will certainly be the case when c_T is large enough, since U is bounded.
[11] That is, $B(p) \geq T(p)$ for any p such that $V < \max\{B(p), T(p)\}$.

McCall 1979*a*; Wilde 1980; and Hey and McKenna 1981). However, many of the important qualitative implications of the model emerge more sharply in the labor market than in the product market.

The labor market analogue concerns an unemployed worker searching for a job. This individual pays a search cost in order to sample vacancies, but only the wage rate is observed. Nonwage characteristics can be observed either by paying an inspection cost or by taking the job.

Two aspects of the labor market analogue are of primary interest—unemployment and the quit rate. Unemployment is related to the expected duration of search. The quit rate is related to the probability that a job which is accepted will subsequently be rejected.

Suppose that the distributions of net utility associated with jobs, conditional on the wage rate, are stochastically decreasing in the wage rate. Then, in the most interesting case, there will be a range of low wages for which renewing search is optimal, a range of intermediate wages for which inspection is optimal, and a range of high wages for which evaluation is optimal (see Sec. IIIB above).

The effects of an increase in information costs on the duration of search seem straightforward. An increase in the cost of search makes both inspection and evaluation more desirable alternatives. Hence the duration of search should fall. An increase in the cost of inspection will likely have ambiguous effects on the duration of search because it makes evaluation a more desirable alternative, but it also makes inspection a less desirable alternative.

The effects of an increase in information costs on the quit rate are less obvious. The argument would seem to go as follows: A quit requires that two events occur: (1) a wage rate is observed for which it is optimal to take the job without observing its nonwage component first and (2) the nonwage component turns out to be too low so that it is optimal to quit and renew search once it is observed. But an increase in either search costs or inspection costs makes evaluation optimal for more wages. In particular, there are *lower* wages for which evaluation is now optimal. These jobs must have higher nonwage components in order to be acceptable; that is, for them the probability of turnover is higher. Hence, the overall probability of a quit should increase as either cost increases.

Unfortunately, these heuristics are incomplete because neither considers the fact that an increase in either c_S or c_T will decrease the return to search, making lower values of the nonwage component acceptable at any given wage. While this reinforces the argument regarding the relationship between information costs and the duration of search, it weakens the argument regarding the relationship

between information costs and the quit rate. In fact, an increase in either cost could either increase or decrease the quit rate.[12]

To justify these assertions formally the model of Section III must be used. Consider first the duration of search. There are really two measures which are of interest. First, $E_S + I_S$ gives the probability that a random price-quality observation will ultimately be acceptable (regardless of how quality is observed). Second, $E_F + E_S + I_S$ gives the probability that a random price observation will cause search to cease (although perhaps only temporarily). It was stated in Section III that

$$\frac{\partial(E_S + I_S)}{\partial c_S} > 0 \quad \text{and} \quad \frac{\partial(E_S + I_S)}{\partial c_T} \gtreqless 0.$$

It can also be shown that

$$\frac{\partial(E_F + E_S + I_S)}{\partial c_S} > 0 \quad \text{and} \quad \frac{\partial(E_F + E_S + I_S)}{\partial c_T} \gtreqless 0.$$

Purchases of goods for which quality has not been observed are really part of the search process. Hence $1/(E_S + I_S)$, which might be called the "pure duration of search," is the proper expression for the expected duration of search. However, empirically it would often be impossible to differentiate between observations which fall in region IV and observations which fall in region V. In other words, the observed duration of search would often correspond to $1/(E_F + E_S + I_S)$, which might be called the "effective duration of search."

Similar problems arise with respect to turnover. In fact, there are three measures of turnover embedded in this model, one ex ante and two ex post. The ex ante measure is simply E_F; it gives the probability that a random price-quality observation will be purchased once and only once. One ex post measure is what might be called the "pure failure rate for evaluation," $\hat{E}_F = E_F/(E_F + E_S)$; it gives the conditional probability that a good will be rejected given that it is purchased without quality having been observed. However, empirically it would often be hard to differentiate between observations which fall in region III and observations which fall in region V. Hence the other ex post measure is what might be called the "effective failure rate for evaluation," $\hat{\hat{E}}_F = E_F/(E_F + E_S + I_S)$; it gives the conditional probability that a good will be rejected given that it is purchased (regardless of whether quality is observed before purchase or after purchase). It turns out that none of these measures is systematically related to either search costs or inspection costs.

[12] See Lippman and McCall (1979*b*) for an extensive discussion of these points in related model.

The third application of this theoretical framework is to the marriage market. The marriage market analogue concerns an unwed individual searching for a marriage partner. This individual pays a search cost in order to sample potential partners, but only some characteristics are observed. Other characteristics can be observed either by paying an inspection cost or by getting married.

The aspect of the marriage market analogue which is of primary interest is dissolution. The most complete analysis of the relationship between information costs and probability of dissolution is provided by Becker, Landes, and Michael (1977). These authors consider two cases, one in which remarriage is impossible and one in which the remarriage market is identical to the marriage market.

Consider first the case in which remarriage is impossible. When remarriage is impossible, the value of dissolution is a constant. In the model analyzed in this paper an analogous assumption is that the value of *not* repurchasing a good is a constant, say $\bar{V}$. Then instead of (3), $B(p)$ would be defined as

$$B(p) = E[U(p, Q) + \beta\bar{k}(p), \tag{3'}$$

where

$$\bar{k}(p) = \bar{V}G[\bar{q}^*(p) \mid p] + \int_{\bar{q}^*(p)}^{\infty} \frac{U(p, q)}{1 - \beta} g(q \mid p)dq,$$

and $\bar{q}^*(p)$ is defined by $U[p, \bar{q}^*(p)] = \bar{V}(1 - \beta)$. The definition of $T(p)$ would remain as in (4). This modification affects the comparative statics of the model in a straightforward way; for $p \leq p_{hT}^*$, $d\bar{q}^*(p)/dc_S = 0 = d\bar{q}^*(p)/dc_T$. Hence as before, $\partial E_S/\partial c_S > 0$ and $\partial E_S/\partial c_T > 0$, but now $\partial E_F/\partial c_S > 0$ and $\partial E_F/\partial c_T > 0$. Furthermore, as before, $\partial \hat{E}_S/\partial c_S < 0$ and $\partial \hat{E}_F/\partial c_T > 0$, but now $\partial \hat{E}_S/\partial c_T < 0$ and $\partial \hat{E}_F/\partial c_T > 0$. However, it remains true, even when the value of *not* repurchasing a good is constant, that $\hat{\hat{E}}_S$ and $\hat{\hat{E}}_F$ bear no systematic relationship to information costs.

These results are, on the surface, consistent with those of Becker, Landes, and Michael. These authors assert that because "the probability of entering into a mismatch" would be greater, "an increase in either the cost of intensive or extensive search would increase the probability of dissolution" (1977, p. 1150).[13] The definition of "probability of dissolution" which these authors use to arrive at this theoretical conclusion is apparently $\hat{E}_F$. The problem is that $\hat{E}_F$ is unobservable (it is $\hat{\hat{E}}_F$ which is observed), and even when remarriage is

[13] Intensive search corresponds to inspection and extensive search corresponds to drawing another observation from ϕ. Hence the cost of intensive search is c_I and the cost of extensive search is c_S.

impossible, there is no systematic relationship between $\hat{E}_F$ and either c_S or c_T!

The situation is even more difficult when the remarriage market is identical to the marriage market. Here, just as in the labor market analogue, none of the partial derivatives relating information costs to turnover can be signed.

These are, of course, somewhat negative observations. In addition, some positive conclusions can be drawn from this analysis. It has been shown that an increase in search costs always decreases the duration of search. An increase in inspection costs will generally increase the duration of search if the effect on $q^*(p)$ is small relative to the effect on p^*_{BT} or p^*_T. This will be the case when the variation in utility due to quality is small relative to the variation in utility due to price.

There appear to be no general conclusions to be drawn regarding the relationship between search costs and turnover. However, as above, an increase in inspection costs will generally increase turnover if the effect on $q^*(p)$ is small relative to the effect on p^*_{BT} or p^*_T. This is again the case when the variation in utility due to quality is small relative to the variation in utility due to price.

V. Conclusion

This paper has established a number of strong results which go against the grain of the extant literature. These results obtain because goods are viewed as multicharacteristic composites in which individual characteristics have specific informational properties.

A number of extensions are possible. An obvious one is to consider goods which are described by more than one nonprice attribute. In this case the consumer would need to decide which attributes to observe before purchase and which to observe after purchase. At each stage, reservation levels would be defined as functions of the actual levels of those attributes already observed. Finally, the consumer would also need to decide in which order to observe attributes. A model of this nature would have important applications in psychology and consumer research, as well as economics. However, one might argue that actual consumers do not use optimizing strategies as complicated as this model would require. Rather, consumers would use various satisficing strategies. In fact, one might apply the same argument to the model analyzed in this paper.

There is much to be said for this argument, but satisficing strategies are even less well understood than optimizing strategies. Furthermore, most empirical work on the duration of search and turnover is based, implicitly or explicitly, on simpler versions of the optimizing model developed in this paper (e.g., Becker, Landes, and Michael

1977). In any event, the issue offers a number of intriguing topics for future research.

References

Becker, Gary S.; Landes, Elisabeth M.; and Michael, Robert T. "An Economic Analysis of Marital Instability." *J.P.E.* 85, no. 6 (December 1977): 1141–88.

Hey, John D., and McKenna, Chris J. "Consumer Search with Uncertain Product Quality." *J.P.E.* 89, no. 1 (February 1981): 54–66.

Lippman, Steven A., and McCall, John J. "The Economics of Belated Information." Working Paper no. 286, Univ. California, Los Angeles, Western Management Sci. Inst., 1979. (*a*)

———. "Search Unemployment: Mismatches, Layoffs, and Unemployment Insurance." Working Paper no. 297, Univ. California, Los Angeles, Western Management Sci. Inst., 1979. (*b*)

MacQueen, J., and Miller, R. G., Jr. "Optimal Persistence Policies." *Operations Res.* 8 (May/June 1960): 362–80.

Nelson, Phillip. "Information and Consumer Behavior." *J.P.E.* 78, no. 2 (March/April 1970): 311–29.

Wilde, Louis L. "An Information-Theoretic Approach to Job Quits." In *Studies in the Economics of Search,* edited by Steven A. Lippman and John J. McCall. Amsterdam: North-Holland, 1979.

———. "On the Formal Theory of Inspection and Evaluation in Product Markets." *Econometrica* 48 (July 1980): 1265–79.

Macroeconomic Policy, Exchange-Rate Dynamics, and Optimal Asset Accumulation

Maurice Obstfeld

Columbia University

The paper develops a model of exchange-rate and current-account determination for a small economy peopled by infinitely lived, utility-maximizing households. In this setting, a central-bank purchase of foreign exchange has no real effects when central-bank foreign reserves earn interest at the world rate and the proceeds are returned to the public. In contrast, an increase in the monetary growth rate does have real effects, even in the long run. The model developed here implies that an increase in government spending may lead to a surplus on current account. The external adjustment process predicted by the model is one in which consumption, real balances, and external assets all rise or fall simultaneously.

Introduction

This paper studies exchange-rate determination and the external adjustment process in a small economy peopled by infinitely lived, utility-maximizing households. The setting is similar to one proposed by Kouri (1976), in that agents are assumed to consume a single good, to hold domestic money, and to have access to a world market in consumption loans. But the approach departs from those prevalent in

Guillermo Calvo suggested the line of investigation pursued in this paper. I am grateful for his comments and for those of Ralph Braid, Robert Cumby, Ronald Findlay, Duncan Foley, Jacob Frenkel, Michael Mussa, and two anonymous referees. Valuable suggestions were made by participants in seminars at Columbia University, the Federal Reserve Board, the University of Chicago, Stanford University Graduate School of Business, the University of Pennsylvania, and the NBER's 1980 Summer Institute in International Studies. All errors are mine. Financial support from the National Science Foundation is gratefully acknowledged.

[*Journal of Political Economy*, 1981, vol. 89, no. 6]
© 1981 by The University of Chicago 0022-3808/81/8906-0009$01.50

recent balance-of-payments models by basing saving behavior and money demand on explicit, intertemporal utility maximization.[1]

The analytical framework is adapted from the seminal work of Sidrauski (1967) and employs the perfect-foresight equilibrium concept explored by Brock (1974), Calvo (1979*a*), and Fischer (1979). In contrast to these contributions, however, the model developed below allows agents' subjective rate of time preference to be endogenously determined in a manner suggested by Uzawa (1968) rather than fixed.[2] This modification permits the small economy we study to attain a stationary long-run equilibrium under conditions of perfect capital mobility.[3]

The plan of the paper, and its main results, are as follows. Section I describes the structure of the economy and the typical household's maximization problem. Particular attention is paid to the budget constraints of the central bank and government. Section II closes the model by imposing continuous money-market equilibrium and endowing agents with perfect foresight concerning the future paths of the price level and government transfer payments. The resulting dynamic system possesses a unique stationary position and a single convergent path, along which consumption, real money balances, and the stock of foreign claims all rise or fall simultaneously. This convergent saddlepath is an equilibrium path and is the focus of the analysis that follows.

Section III analyzes the effects of some unanticipated macroeconomic policy actions. A central-bank purchase of foreign exchange is found to have no real effects when official foreign reserves earn interest that is distributed to the public. In contrast, an increase in the monetary growth rate does have real effects: It occasions exchange-rate overshooting (in Flood's [1979] sense) and a current-account surplus and leads, in the long run, to higher levels of consumption and foreign claims. Finally, the model implies that an increase in government spending may induce a surplus on current account, contrary to the usual presumption. A deficit is possible when government spending produces a public good having a high marginal value to consumers. The explicit specification of agents' objective

[1] Research along similar lines is reported in papers by Calvo (1979*b*, 1980), Helpman (1979), Razin (1979), and Stockman (1980). Obstfeld (1980) describes a two-country extension of this paper's model.

[2] Endogenous time preference has been discussed in the trade literature in papers by Calvo and Findlay (1978) and by Findlay (1978). Kouri (1980) adopts a consumption function suggested by Uzawa's (1968) theory. The present approach is applied in a somewhat different context in Obstfeld (1981).

[3] This could be accomplished also by allowing bonds (in addition to real balances) to enter consumers' utility functions, as in Calvo (1980).

functions is particularly useful in Section III, for it permits a natural welfare analysis of alternative macroeconomic policies.

Section IV contains concluding remarks. Appendices supply the technical details justifying some of the paper's assertions.

I. The Model

We consider an open economy inhabited by identical households, each maximizing its welfare over an infinite lifetime. As in Kouri (1976), there is a single, perishable, consumption good, and households' financial wealth is divided between domestic fiat money and internationally traded bonds denominated in foreign currency. The economy is small, and it can therefore influence neither the foreign currency price of the consumption good, P^*, nor the world bond rate, r, both of which are assumed constant. The link between the domestic price level, P, and P^* is provided by the relationship $P = EP^*$, where E denotes the domestic money price of foreign exchange. This exchange rate is allowed to float freely by the monetary authority. On the assumption that foreigners do not hold domestic money, E adjusts to maintain equality between the real money supply and domestic real money demand at each instant. We adopt the normalization $P^* = 1$.

The number of households in the economy is, for convenience, also taken to be 1. The household's instantaneous utility is a separable function $U(c,m) = u(c) + v(m)$ of consumption, c, and real money holdings, m, defined as nominal money holdings, M, deflated by the home price level, $m \equiv M/P$. The functions $u(\cdot)$ and $v(\cdot)$ are nonnegative, strictly concave, and twice continuously differentiable, with the property that

$$\lim_{c\to 0} u'(c) = \lim_{m\to 0} v'(m) = \infty. \tag{1}$$

Assumption (1) eliminates the possibility of corner solutions. It is clear, in addition, that the derivatives $u'(\cdot)$ and $v'(\cdot)$ are everywhere strictly positive.

The household seeks to maximize the discounted sum of future instantaneous utilities,

$$V = \int_0^\infty U(c_t,m_t)e^{-\Delta_t}dt. \tag{2}$$

The discount factor Δ_t is defined by

$$\Delta_t \equiv \int_0^t \delta_s ds, \tag{3}$$

where δ_s is the instantaneous subjective discount rate at time s.[4] Following Uzawa (1968), we postulate that δ_s is a function of utility at time s,

$$\delta_s = \delta[U(c_s, m_s)], \tag{4}$$

and that the function $\delta(\cdot)$ is positive and satisfies

$$\delta'(U) > 0, \delta''(U) > 0, \delta(U) - U\delta'(U) > 0. \tag{5}$$

The first two conditions listed in (5) are in no way implied by the hypothesis of endogenous time preference but are imposed because they are analytically convenient.[5] The last of the conditions ensures that the present discounted value $\bar{U}/\delta(\bar{U})$ of a stationary utility stream $\bar{U}$ is increasing in $\bar{U}$.

At each moment t, the household must allocate its income between current expenditure and the accumulation of real wealth. Real output, y, is taken to be exogenous and fixed. The other components of expected disposable income are interest payments from abroad, equal to the world bond rate, r, times the real value F_t of the family's net foreign claims; expected real transfers from the government, τ_t; and expected capital gains on real balances, equal to $-\pi_t m_t$, where π_t denotes the expected inflation rate. Defining the household's real assets at time t (excluding the capitalized value of expected future transfers) as

$$a_t \equiv \frac{y}{r} + m_t + F_t, \tag{6}$$

the flow constraint linking asset accumulation to saving may be written as

$$\dot{a}_t = ra_t + \tau_t - c_t - (\pi_t + r)m_t, \tag{7}$$

while the intertemporal budget constraint takes the form

$$\int_0^\infty e^{-rt}[c_t + (\pi_t + r)m_t]dt \leq a_0 + \int_0^\infty e^{-rt}\tau_t dt. \tag{8}$$

Restriction (8) is important in the present setting of perfect capital mobility, for it rules out the possibility that the household can attain unbounded utility by borrowing arbitrarily large sums in the world

[4] When $\delta_s = \delta$, a constant, for all s, we have the usual discount factor, $\Delta_t = \delta t$

[5] The nature of this convenience becomes apparent in applying the maximum principle to solve the household's lifetime problem and in proving the existence of a stable perfect-foresight equilibrium path converging to the economy's stationary state. These points are taken up again in context. A general discussion of the time preference concept is found in Koopmans, Diamond, and Williamson (1964).

capital market and meeting all interest payments through further borrowing. As Appendix B shows, (7) and (8) confine the household to paths along which

$$a_t + \int_t^{\infty} e^{-r(s-t)}\tau_s ds \geq 0$$

for all t, so that household net worth is always nonnegative.

The household's problem is to find $\{c_t, m_t\}$ maximizing lifetime welfare V subject to the constraints (3), (7), and (8), given an initial level a_0 of real assets. The calculations are simplified by changing variables from t to Δ ("psychological time") using the fact, implied by (3), that $d\Delta = \delta[U(c_t,m_t)]dt$. The household's problem may then be expressed as that of choosing time paths for consumption and real balances that satisfy (8) and[6]

$$\text{maximize} \int_0^{\infty} \frac{U(c,m)}{\delta[U(c,m)]} e^{-\Delta} d\Delta$$

$$\text{subject to } \frac{da}{d\Delta} = \frac{ra + \tau - c - (\pi + r)m}{\delta[U(c,m)]}.$$

Necessary conditions for an optimal program are given by the maximum principle.[7] These require that c and m be chosen so that for each value of the discount factor Δ, the Hamiltonian

$$H = \frac{U(c,m)}{\delta[U(c,m)]} + \lambda \frac{[ra + \tau - c - (\pi + r)m]}{\delta[U(c,m)]} \tag{9}$$

is maximized. In (9), $\lambda = \lambda_\Delta$ is the co-state variable, which is interpreted as the shadow value, in utility terms, of real assets. The implied first-order conditions (at the interior maximum) are

$$(1 - (\delta'/\delta)\{U + \lambda[ra + \tau - c - (\pi + r)m]\})U_c = \lambda, \tag{10}$$

$$(1 - (\delta'/\delta)\{U + \lambda[ra + \tau - c - (\pi + r)m]\})U_m = \lambda(\pi + r). \tag{11}$$

An additional necessary condition is that the shadow price of assets evolves according to the law,

$$\frac{d\lambda}{d\Delta} = \lambda\left(\frac{\delta - r}{\delta}\right). \tag{12}$$

Equations (10) and (11) imply the usual necessary condition of static utility maximization,

[6] The psychological time subscript, Δ, is omitted for the sake of notational simplicity.

[7] See Arrow and Kurz 1970. If $U(c,m)$ is strictly concave and both consumption and real balances are normal goods, the conditions listed in (5) imply that the maximand $U(c,m)/\delta[U(c,m)]$ is strictly concave.

$$x(c_t,m_t) \equiv U_m(c_t,m_t)/U_c(c_t,m_t) = \pi_t + r, \tag{13}$$

while (12) implies that the time derivative of λ_Δ is given by[8]

$$\dot{\lambda}_t = \lambda_t\{\delta[U(c_t,m_t)] - r\}. \tag{14}$$

It remains to specify the behavior of the government and the central bank. We consolidate the balance sheets of these two entities and work with a single public-sector budget constraint. A key assumption is that the stock of central-bank foreign reserves earns interest at the world rate r.

The public sector consumes goods and makes (net) transfers, financing any deficit through money creation. Thus, all government debt is purchased by the central bank, none by private citizens. Letting g denote the (constant) level of real government consumption per household, the public sector's budget constraint is $g = \dot{M}_t/P_t + rR - \tau_t = (\dot{M}_t/M_t)m_t + rR - \tau_t$, where R denotes the (constant) level of central-bank foreign reserves, measured in terms of output, and τ_t, as before, denotes net real transfers.

We shall assume that the central bank varies τ_t in such a way as to hold the rate of monetary growth $\dot{M}_t/M_t$ constant at μ.[9] The level of transfers is then determined according to the formula

$$\tau_t = \mu m_t + rR - g. \tag{15}$$

A final assumption is that μ is always chosen so that $\mu + r > 0$. This inequality guarantees the existence of a stationary state.

II. Perfect-Foresight Equilibrium Dynamics

The model of the previous section is closed by the assumption that the path of the economy is a perfect-foresight equilibrium path, in the sense of Brock (1974). Let $\{\hat{P}_t\}$ be a (differentiable) price-level path and $\{\widehat{P_t\tau_t}\} = \{\mu e^{\mu t}M_0 + \hat{P}_t rR - \hat{P}_t g\}$ the associated path of nominal transfer payments from the government, where M_0 denotes the nominal money stock at $t = 0$. Acting on the belief that these paths of the price level and nominal transfers will prevail, the household can calculate expected real transfers $\{\hat{\tau}_t\}$ and determine optimal paths for

[8] Only necessary conditions (10) and (11) differ from what they would be in the standard intertemporal optimization framework, where $\delta_t = \delta$, an exogenous constant. In the standard setup, (10) is replaced by $U_c = \lambda$. But in the present context, an increase in today's consumption affects not only today's instantaneous utility but also the discount factor applied to all utilities enjoyed in the future. Note also that if the instantaneous subjective discount rate δ were constant, necessary condition (14) would rule out the possibility of a stationary state unless it so happened that δ and r were equal.

[9] This assumption is crucial for the foreign-exchange intervention "neutrality" proposition derived below and differs from the one made by Kouri (1976), who takes government consumption to be the residual item.

its consumption rate and real balances, subject to (7) and (8) with initial real assets $a_0 = (y/r) + F_0 + (M_0/\hat{P}_0)$. The path $\{\hat{P}_t\}$ is a perfect-foresight equilibrium path if the desired path of real balances, $\{\hat{m}_t\}$, equals the actual path, that is, if

$$\hat{m}_t = M_t/\hat{P}_t = e^{\mu t}M_0/\hat{P}^t \tag{16}$$

for all $t \geqslant 0$.

The necessary conditions derived in the previous section may be used to find the collection of price-level paths $\{P_t\}$ consistent with perfect foresight and individual optimality. The differential equations (7) and (14), together with the conditions (6), (10), (13), (16), and

$$\pi_t = \dot{P}_t/P_t, \tag{17}$$

must be satisfied by any perfect-foresight equilibrium trajectory of the economy. We now investigate the restrictions these requirements place on the possible equilibrium paths of bonds F_t, consumption c_t, real balances m_t, and, by implication, the home price level. Recalling our earlier convention that the world price level $P^* = 1$, we shall identify the path of the domestic price level with that of the exchange rate, E.

Differentiating equilibrium condition (16), we find that desired real balances must satisfy the relation

$$\dot{P}_t/P_t = \mu - (\dot{m}_t/m_t) \tag{18}$$

along an equilibrium path. Combining (17) and (18) with necessary condition (13) yields

$$\dot{m}_t = [\mu + r - x(c_t,m_t)]m_t. \tag{19}$$

The differential equation (19) must govern the evolution of real balances in perfect-foresight equilibrium.

The flow constraint (7) may be written in the form $\dot{m}_t + \dot{F}_t = y + rF_t + \tau_t - c_t - \pi_t m_t$. When we use (17), (18), and the public-sector budget constraint (15), this differential equation becomes

$$\dot{F}_t = y + r(F_t + R) - c_t - g, \tag{20}$$

the familiar relation between the current and capital accounts under a floating exchange rate. Equation (20), which equates the home country's foreign lending to the excess of domestic income over absorption by the domestic private and public sectors, describes the equilibrium path of foreign bond holdings. Note that because foreigners do not hold domestic money, the variable F_t is predetermined by the past history of the current account. It can jump instantaneously only through central-bank purchases and sales and must otherwise adjust slowly over time.

It remains to derive the differential equation governing movements in consumption. Using the necessary condition (10), we may solve for the co-state variable λ_t with the help of (15) and (17)–(19):

$$\lambda_t = \lambda(c_t, m_t, F_t + R) =$$

$$\frac{[\delta(U_t) - U_t\delta'(U_t)]u'(c_t)}{\delta(U_t) + \{y + r(F_t + R) + [\mu + r - x(c_t, m_t)]m_t - c_t - g\}\delta'(U_t)u'(c_t)}.$$

Differentiating λ with respect to time and applying (14) yields $\dot{\lambda}_t = \lambda_c\dot{c}_t + \lambda_m\dot{m}_t + \lambda_F\dot{F}_t = \lambda_t\{\delta[U(c_t, m_t)] - r\}$, which may be solved for the time derivative of consumption,

$$\dot{c}_t = \lambda_c^{-1}(\lambda_t\{\delta[U(c_t, m_t)] - r\} - \lambda_m\dot{m}_t - \lambda_F\dot{F}_t) \equiv \psi(c_t, m_t, F_t + R). \quad (21)$$

Equation (21) is the final differential equation of a three-equation system.[10] Together, equations (19)–(21) must describe the evolution of the economy along any perfect-foresight equilibrium path.

Not every path satisfying (19)–(21) is a perfect-foresight equilibrium path, however, for the conditions from which these paths are derived are necessary for optimality, but are not, in general, sufficient. For the balance of this paper, we shall focus attention on paths that bring the economy to the *stationary state* $(\bar{c}, \bar{m}, \bar{F})$, defined by the conditions $\psi(\bar{c}, \bar{m}, \bar{F} + R) = \mu + r - x(\bar{c}, \bar{m}) = y + r(\bar{F} + R) - \bar{c} - g = 0$. As is easily verified, the stationary state exists and is unique when, as we have assumed, $\mu + r > 0$.

Appendix A proves that the system (19)–(21), when linearized in a neighborhood of the stationary state, possesses a single negative characteristic root, and so a unique convergent path. Appendix B proves that this path is a perfect-foresight equilibrium path: The choices of real balances and consumption it dictates are optimal choices from the household's standpoint, given the implied initial value of real assets and the implied anticipated paths of prices and transfers. We therefore assume that for any value of the predetermined stock of foreign claims, the equilibrium price level and consumption rate are those placing the economy on the path converging to the stationary or long-run equilibrium.[11] It should be noted that this convergent saddlepath satisfies the household budget constraint (8), for, along it,

$$\int_0^\infty e^{-rt}[c_t + (\pi_t + r)m_t - \tau_t]dt = a_0 - \lim_{t\to\infty} a_t e^{-rt} = a_0. \quad (22)$$

To find the saddlepath, we let θ_1 denote the system's negative characteristic root and $\vec{\omega} = [\omega_{11}, \omega_{21}, \omega_{31}]'$, an eigenvector belonging to

[10] The equation giving $\dot{c}$ cannot be defined if $\lambda_c = 0$. It is shown in Appendix A that $\lambda_c < 0$ in a neighborhood of the system's stationary state. In fact, it can be shown that $\lambda_c < 0$ globally if hyperdeflations are ruled out.

[11] It has not been demonstrated that perfect-foresight equilibrium is locally unique.

θ_1. Any *convergent* solution of the linearized system must take the form

$$c_t - \bar{c} = \omega_{11} k \exp(\theta_1 t),$$
$$m_t - \bar{m} = \omega_{21} k \exp(\theta_1 t),$$
$$F_t - \bar{F} = \omega_{31} k \exp(\theta_1 t),$$

where k is an arbitrary constant determined by the value of F at time $t = 0$.[12] Differentiating, we find that

$$\dot{c}_t = \theta_1(c_t - \bar{c}) = \theta_1(\omega_{11}/\omega_{31})(F_t - \bar{F}), \tag{23}$$

$$\dot{m}_t = \theta_1(m_t - \bar{m}) = \theta_1(\omega_{21}/\omega_{31})(F_t - \bar{F}), \tag{24}$$

$$\dot{F}_t = \theta_1(F_t - \bar{F}), \tag{25}$$

when the economy approaches the stationary state. The saddlepath of the system is simply the intersection of the two hyperplanes defined by equations (23) and (24).

The importance of equations (23)–(25) is that they allow us to characterize the motion of consumption, real balances, and external asset holdings as the economy converges to long-run equilibrium. As is shown in Appendix A, the components of the eigenvector $\vec{\omega}$ must all be of the same sign, and an immediate consequence is that the three variables must always rise or fall together along the saddlepath.[13] This is a key property, for it allows us to trace the path taken by the economy in response to various policy surprises. However, one particular implication of this result is important enough to be mentioned before we turn to such exercises. When the current account is in surplus ($\dot{F}_t > 0$), real balances must be rising, and when it is in deficit ($\dot{F}_t < 0$), they must be falling. Thus, the exchange rate's depreciation rate will fall short of the rate of monetary growth when there is an external surplus and will exceed it when there is a deficit. This fundamental relationship between exchange-rate depreciation and the current account appears in models developed by Kouri (1976), Calvo and Rodríguez (1977), and others. But it emerges here from the hypothesis of optimizing behavior and not as a consequence of the assumed properties of consumption and money-demand functions.

Figure 1 shows how the economy's long-run equilibrium is found. From (14), stationary-state utility $\bar{U}$ is determined by the equality of the marginal rate of time preference and the world bond rate,[14]

[12] See Hirsch and Smale 1974.

[13] Calvo (1979*a*) provides a similar characterization of the saddlepath arising in the Brock-Sidrauski perfect-foresight model.

[14] From (10), $\bar{\lambda} = \{[\delta(\bar{U}) - \bar{U}\delta'(\bar{U})]/\delta(\bar{U})\}u'(\bar{c})$. The last inequality in (5) therefore implies that $\bar{\lambda}$ is strictly positive. This justifies condition (26).

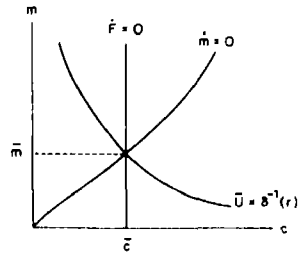


FIG. 1

$$\delta(\bar{U}) = r, \tag{26}$$

and so, stationary-state consumption $\bar{c}$ and real balances $\bar{m}$ must satisfy $U(\bar{c},\bar{m}) = \bar{U}$, as shown in the diagram. The second requirement on the point $(\bar{c},\bar{m})$ is that it lie on the locus of points consistent with $\dot{m} = 0$; the latter is, by (19), the Engel curve associated with an opportunity cost $\mu + r$ of holding real balances. It slopes upward from the origin. Given the value of $\bar{c}$ determined by $\bar{U}$ and $\mu + r$, $\bar{F}$ is determined by the relation $\bar{F} = [(\bar{c} + g - y)/r] - R$. The locus labeled $\dot{F} = 0$ in figure 1 is vertical at the unique private consumption level consistent with current-account balance given stationary-state bond holdings $\bar{F}$. An increase in $\bar{F}$ would increase long-run domestic income and so shift this schedule rightward.

III. The Effects of Macroeconomic Disturbances

This section studies the impact and long-run effects of intervention in the foreign exchange market, an increase in the rate of monetary expansion, and an increase in real government consumption. The macroeconomic disturbances are assumed to take the public by surprise, but they are permanent and lead to no expectation of future policy actions. The economy's initial position is the stationary state.

Foreign Exchange Intervention

The central bank intervenes in the foreign exchange market by purchasing foreign claims from the public with domestic money. The transaction accomplishes a transfer of interest-earning assets from the public to the central bank as well as an increase in the nominal money supply. But the *sum* $\bar{F} + R$ of privately owned bonds and central-bank reserves is unaffected by this transfer.

Now, equations (20) and (21) involve only the economy's aggregate claims on the rest of the world; the distribution of their ownership

between the public and the central bank is irrelevant. Thus, the long-run value of F simply declines from $\bar{F}$ to $\bar{F} - \Delta R$ as a result of the intervention, while $\bar{c}$ and $\bar{m}$ are unchanged. Because the stock of privately owned foreign assets is immediately at its new long-run level, consumption and real balances must, by (23) and (24), remain at their respective stationary-state levels $\bar{c}$ and $\bar{m}$. The monetary authority's action occasions a rise in the price level (a depreciation of the exchange rate) exactly proportional to the increase in the nominal supply of money. It has no real effects. Money creation accomplished through a purchase of foreign exchange has the same impact as a "helicopter" money-supply change of equal magnitude.

It is sometimes asserted that there *is* a difference between these two financial policies, in that an official purchase leads to a current surplus as domestic residents seek to restore their holdings of external claims to the original level.[15] The difference disappears here because foreign assets held by the central bank continue to earn interest at the rate r, while households capitalize all transfers from the government (see Mundell 1971, chap. 1). By the government budget constraint (15), the stream of interest earnings taken away from the public through central-bank intervention is returned in the form of transfers. Households therefore suffer no fall in permanent disposable income when a portion of their bond holdings is transferred to the central bank.

This neutrality result would clearly break down if, for example, reserves were held in some barren, non-interest-bearing form, or if increased central-bank earnings were used to augment government consumption rather than transfer payments. That intervention has real effects in these circumstances is to be expected, for it involves real fiscal measures in addition to a neutral change in the nominal supply of money.

An Increase in the Monetary Growth Rate

Figure 2 depicts the effects of an increase in the rate of monetary growth, μ. As (26) shows, the long-run instantaneous utility level of the representative household is unchanged by the disturbance. But as the rate of monetary growth and the inflation rate coincide in the long run, the disturbance raises the opportunity cost of holding money in the stationary state. Thus, the household economizes on real balances and consumes more in the new long-run equilibrium. This is represented in figure 2 by the downward shift of the Engel curve along

[15] See, e.g., Kouri 1978.

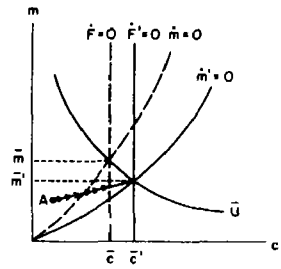


FIG. 2

which real balances are stationary. The rise in long-run consumption entails a rise in long-run bond holdings. Accordingly, the $\dot{F} = 0$ locus shifts to the right, so that it passes through the new asymptotic equilibrium point, $(\bar{c}', \bar{m}')$.

Knowledge of the long-run effects of the policy shift allows us to infer its short-run effects using the properties of the convergent saddlepath described in Section II. Because the eventual level of external assets is higher after the rise in μ, consumption must initially decline to bring about the implied current-account surplus. And because real balances rise during the course of a surplus and $\bar{m}'$ is lower than $\bar{m}$, the policy also entails a sharp initial fall in real balances and so a depreciation of the exchange rate. The economy's short-run, temporary equilibrium at $t = 0$ is thus at a point like A in figure 2.

Both consumption and real balances rise from the levels given by point A as the stock of foreign claims approaches its higher, asymptotic level. This implies, in particular, that the rate of depreciation of the exchange rate during the transition will fall short of the higher rate of monetary expansion. The exchange rate thus exhibits an "overshooting" pattern of adjustment, in that the impact fall in real balances exaggerates the long-run decline. This reproduces, in an optimizing framework, the result obtained by Kouri (1976), Calvo and Rodríguez (1977), and Flood (1979) on the basis of assumed aggregate relationships.

It is of interest to note that the integral of discounted utilities given in equation (2) is lower along the transition path following a rise in μ than it would have been along the original, unperturbed stationary-state path. This follows from the last inequality in (5) and the observation that the household's instantaneous utility falls the moment μ increases, rising to its original level only in the long run. Thus, an increase in the growth rate of money, in the present setting, has an unambiguously negative effect on the welfare of economic agents.

An Increase in Government Consumption

We analyze, finally, the consequences of an increase Δg in real government consumption. It is assumed initially that government consumption is wasteful, in that it does not enter into the household's utility function. An example in which government spending produces a public good complementary to private consumption is analyzed next.

In the absence of any change in private consumption from its initial level $\bar{c}$, an increase in government consumption would entail a current-account deficit at the initial level, $\bar{F}$, of foreign assets. This is shown in figure 3 by the leftward shift of the $\dot{F} = 0$ locus to $\dot{F}' = 0$, which is vertical at the lower level of private consumption compatible with external balance the moment after the increase in g occurs.

The steady-state levels of consumption and real balances are not altered by the government's action, however, for these variables are determined, independently of the level of g, by the two conditions, $\delta[U(\bar{c},\bar{m})] = r$, $x(\bar{c},\bar{m}) = \mu + r$. It follows that the stock of external claims associated with the economy's new stationary state must be higher than the initial stock, $\bar{F}$: In the long-run equilibrium following the increase in public-sector expenditure, the private consumption level $\bar{c}$ can be sustained with no net borrowing from abroad only if the economy's privately owned foreign assets rise by the amount $\Delta g/r$, so that additional interest payments from foreigners finance the additional national absorption.

Because the economy's adjustment path is noncyclical, the current account must go into surplus when the rise in g occurs. Consumption thus declines by an amount *greater* than the increase in government spending. And because consumption, real balances, and foreign bond holdings rise together along the saddlepath, the exchange rate necessarily depreciates on impact, reducing real balances in the short run so that they may subsequently increase toward their long-run level, $\bar{m}$.

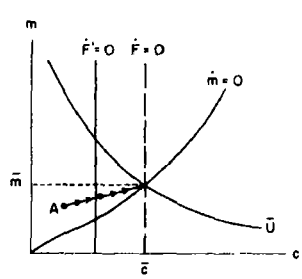


FIG. 3

The point labeled A in figure 3 again represents short-run equilibrium at $t = 0$.

When the monetary growth rate, μ, is zero, the government's action is simply a tax-financed increase in government spending. The model's prediction contradicts the standard Keynesian presumption that the accompanying short-run fall in disposable income must reduce private saving along with private expenditure, pushing the current account into deficit. In the present framework, private saving must actually increase in order that the economy may attain, in the long run, the higher level of national income consistent with the predisturbance private consumption level. Looked at another way, a tax-financed increase in government spending entails a more than complete crowding out of private aggregate demand in the short run.

The preceding discussion has been based on the assumption that the level of government spending does not enter the utility function, as it would if government consumption resulted in the provision of some public good. To illustrate how the analysis might change if this assumption were modified, we study a particular example.

Assume now that the utility function has the form

$$U(c,m,g) = u(c,g) + v(m), u_g, u_{cg} > 0. \tag{27}$$

According to (27), public and private consumption are complementary goods. The stationary values of consumption, real balances, and foreign assets are determined by the equations

$$\delta[U(\bar{c},\bar{m},g)] = r, \tag{28}$$

$$v'(\bar{m})/u_c(\bar{c},g) = \mu + r, \tag{29}$$

$$\bar{F} = [(\bar{c} + g - y)/r] - R. \tag{30}$$

In particular, (28) and (29) imply that $\bar{c}$ and $\bar{m}$ are no longer independent of g, as they were in the case of purely wasteful government spending.

The long-run derivatives of $\bar{c}$, $\bar{m}$, and $\bar{F}$ with respect to g are:

$$\frac{d\bar{c}}{dg} = -rD^{-1}[\bar{u}_g\bar{v}'' + \bar{v}'(\mu + r)\bar{u}_{cg}], \tag{31}$$

$$\frac{d\bar{m}}{dg} = rD^{-1}(\mu + r)(\bar{u}_c\bar{u}_{cg} - \bar{u}_g\bar{u}_{cc}), \tag{32}$$

$$\frac{d\bar{F}}{dg} = D^{-1}[(\bar{u}_c - \bar{u}_g)\bar{v}'' + \bar{v}'(\mu + r)(\bar{u}_{cc} - \bar{u}_{cg})], \tag{33}$$

where $D = r[\bar{u}_c\bar{v}'' + (\mu + r)\bar{u}_{cc}\bar{v}'] < 0$ and functions beneath overbars are evaluated at the initial long-run equilibrium. Only (32) can be

unambiguously signed under the assumption $u_{cg} > 0$.[16] In particular, if the marginal utility of private consumption is sufficiently low relative to that of public consumption, an increase in g may, as (33) shows, occasion a current deficit, contrary to the earlier analysis. In this situation, private consumption falls by an amount smaller than Δg in the short run; however, there is a decline in *long-run* consumption that exceeds Δg.

In the event that the level of public-good provision happens to be initially at the Samuelson static optimum, so that $\bar{u}_c = \bar{u}_g$, (32) and (33) imply that the increase in g must occasion a current surplus and a depreciation, as before. Why must stationary consumption rise above $c - \Delta g$ in this case? If, in the long run, consumption were to fall by exactly Δg with $\bar{m}$ unchanged, the economy would be at its long-run utility level $\bar{U}$, but with $v'(\bar{m})/u_c(\bar{c} - \Delta g, g + \Delta g) < \mu + r$, in violation of (29). Long-run equilibrium thus requires that real money balances fall below $\bar{m}$ and that consumption rise above $\bar{c} - \Delta g$.

The welfare analysis of an increase in government consumption is straightforward. We note here simply that when $u_{cg} > 0$ and $\bar{u}_c = \bar{u}_g$, an increase in g must decrease the discounted sum of future utilities V below $\bar{U}/\delta(\bar{U})$, for it induces a current surplus and so a short-run decline in private consumption greater than Δg. The implication is that a level of public expenditure optimal from a static standpoint may no longer appear optimal once full account is taken of the dynamic repercussions of a change.

IV. Conclusion

This paper has developed an open-economy model based on household utility maximization over an infinite lifetime. The model answers several questions concerning the effects of macroeconomic policies when exchange rates float freely. In some cases, the answers coincide with results derived previously on the basis of more or less mechanistic behavioral assumptions. But in other cases, the answers contradict these results.

In particular, the model implies that money is not superneutral in either the short or long runs and places both the exchange-rate "overshooting" hypothesis and the association between depreciation and current deficits on a firm microeconomic foundation. However, it

[16] The following considerations imply that $d\bar{m}/dg < 0$. Because $u_{cg} > 0$, the increase in g would entail an increase in the marginal utility of private consumption at the initial level of c. If $\bar{m}$ were to rise, v' would fall, and (29) could not be satisfied in the long run in the absence of a higher stationary consumption level. But with $\bar{c}$, $\bar{m}$, and g all higher, the household's long-run utility level would exceed $\bar{U}$, contrary to (28). Thus, $\bar{m}$ declines and $v'(\bar{m})$ rises.

casts doubt on the validity of inferences drawn from the Keynesian consumption function which, in some form, is central to the classic Mundell-Fleming model and most subsequent work. In addition, the approach highlights the importance of the government and central-bank budget constraints in discussions of the effects of financial policy measures.

Appendix A

This appendix provides the proof of local saddlepath stability for the system (19)–(21) and characterizes the unique path leading to stationary-state equilibrium.

The model involves two state variables, c and m, which may take (unanticipated) discrete jumps, and a third state variable, F, which is predetermined by the past history of external asset accumulation. Thus, to prove saddlepath stability, it must be shown that when linearized about its stationary position, the differential equation system,

$$\dot{c} = \frac{\lambda\{\delta[u(c) + v(m)] - r\} - \lambda_m[\mu + r - x(c,m)]m - \lambda_F[y + r(F + R) - c - g]}{\lambda_c}$$

$$= \psi(c,m, F + R),$$

$$\dot{m} = [\mu + r - x(c,m)]m,$$

$$\dot{F} = y + r(F + R) - c - g,$$

possesses two positive characteristic roots and one negative characteristic root, implying the existence of a unique path converging to long-run equilibrium. Recall that

$$\lambda = \frac{\{\delta - [u(c) + v(m)]\delta'\}u'}{\delta + \{y + r(F + R) + [\mu + r - x(c,m)]m - c - g\}\delta'u'},$$

$$x(c,m) \equiv v'(m)/u'(c).$$

At the system's long-run equilibrium $(\bar{c},\bar{m},\bar{F})$, we have $\dot{c} = \dot{m} = F = a = \delta[u(\bar{c}) + v(\bar{m})] - r = 0$, and, linearizing in a neighborhood of this point, the system becomes

$$\dot{c} = \bar{\psi}_c(c - \bar{c}) + \bar{\psi}_m(m - \bar{m}) + \bar{\psi}_F(F - \bar{F}),$$

$$\dot{m} = -\bar{x}_c\bar{m}(c - \bar{c}) - \bar{x}_m\bar{m}(m - \bar{m}),$$

$$F = -(c - \bar{c}) + r(F - \bar{F}).$$

(Functions beneath overbars are evaluated at $[\bar{c},\bar{m},\bar{F}]$.) Using the definition of long-run equilibrium, we find that

$$\bar{\psi}_c = (\bar{\lambda}/\bar{\lambda}_c)\bar{\delta}'\bar{u}' + (\bar{\lambda}_m/\bar{\lambda}_c)\bar{x}_c\bar{m} + (\bar{\lambda}_F/\bar{\lambda}_c), \tag{A1}$$

$$\bar{\psi}_m = (\bar{\lambda}/\bar{\lambda}_c)\bar{\delta}'\bar{v}' + (\bar{\lambda}_m/\bar{\lambda}_c)\bar{x}_m\bar{m}, \tag{A2}$$

$$\bar{\psi}_F = -(\bar{\lambda}_F/\bar{\lambda}_c)r, \tag{A3}$$

where

$$\bar{\lambda}_c = \frac{[\bar{\delta} - (\bar{u} + \bar{v})\bar{\delta}'][\bar{\delta}\bar{u}'' + \bar{\delta}'(\bar{u}')^2\bar{x}_c\bar{m}] - (\bar{u} + \bar{v})(\bar{u}')^2\bar{\delta}\bar{\delta}'}{\bar{\delta}^2}, \tag{A4}$$

$$\bar{\lambda}_m = \frac{[\bar{\delta} - (\bar{u} + \bar{v})\bar{\delta}'](\bar{\delta}'\bar{u}'\bar{x}_m\bar{m} - \bar{\delta}'\bar{v}')\bar{u}' - (\bar{u} + \bar{v})\bar{u}'\bar{v}'\bar{\delta}\bar{\delta}''}{\bar{\delta}^2}, \tag{A5}$$

$$\bar{\lambda}_F = \frac{-[\bar{\delta} - (\bar{u} + \bar{v})\bar{\delta}']\bar{\delta}'(\bar{u}')^2}{\bar{\delta}} < 0. \tag{A6}$$

Now $\bar{x}_c = -v'(\bar{m})u''(\bar{c})/[u'(\bar{c})]^2 > 0$ and $\bar{x}_m = v''(\bar{m})/u'(\bar{c}) < 0$; and since $U \geqslant 0$, $\delta - U\delta' > 0$, and $\delta'' > 0$ by assumption,

$$\bar{\lambda}_m < 0. \tag{A7}$$

Further, substituting for $\bar{x}_c$ in (A4) yields

$$\bar{\lambda}_c = \frac{[\bar{\delta} - (\bar{u} + \bar{v})\bar{\delta}'](\bar{\delta} - \bar{\delta}'\bar{v}'\bar{m})\bar{u}'' - (\bar{u} + \bar{v})(\bar{u}')^2\bar{\delta}\bar{\delta}''}{\bar{\delta}^2}. \tag{A8}$$

Because $u(\cdot)$ and $v(\cdot)$ are nonnegative and concave, we have, using (5), $\bar{\delta} - \bar{\delta}'\bar{v}'\bar{m} > \bar{\delta} - \bar{v}\bar{\delta}' > \bar{\delta} - (\bar{u} + \bar{v})\bar{\delta}' > 0$, and (A8) implies immediately that

$$\bar{\lambda}_c < 0. \tag{A9}$$

We note, finally, that because $\bar{\lambda} = \{\bar{\delta} - [u(\bar{c}) + v(\bar{m})]\bar{\delta}'\}u'/\bar{\delta}$, (A1) and (A6) imply that

$$\bar{\psi}_c = (\bar{\lambda}_m/\bar{\lambda}_c)\bar{x}_c\bar{m} > 0. \tag{A10}$$

Using (A10), the linearized system becomes

$$\begin{bmatrix} \dot{c} \\ \dot{m} \\ \dot{F} \end{bmatrix} = \begin{bmatrix} \dfrac{\bar{\lambda}_m\bar{x}_c\bar{m}}{\bar{\lambda}_c} & \dfrac{\bar{\lambda}\bar{\delta}'\bar{v}' + \bar{\lambda}_m\bar{x}_m\bar{m}}{\bar{\lambda}_c} & \dfrac{-\bar{\lambda}_F r}{\bar{\lambda}_c} \\ -\bar{x}_c\bar{m} & -\bar{x}_m\bar{m} & 0 \\ -1 & 0 & r \end{bmatrix} \begin{bmatrix} c - \bar{c} \\ m - \bar{m} \\ F - \bar{F} \end{bmatrix}. \tag{A11}$$

By (A10) and the fact that $\bar{x}_m < 0$, the trace of the system's matrix is positive. Since this trace is also equal to the sum of the system's characteristic roots, there must be at least one root with positive real part.

On the other hand, we may write the system's determinant as

$$\begin{aligned} \Omega &= \{\bar{x}_m\bar{m}(\bar{\lambda}_F/\bar{\lambda}_c)r\} + r(\bar{x}_c\bar{m})(\bar{\lambda}\bar{\delta}'\bar{v}' + \bar{\lambda}_m\bar{x}_m\bar{m})/\bar{\lambda}_c - r(\bar{x}_m\bar{m})(\bar{\lambda}_m/\bar{\lambda}_c)\bar{x}_c\bar{m} \\ &= \{\bar{x}_m\bar{m}(\bar{\lambda}_F/\bar{\lambda}_c)r\} + r(\bar{x}_c\bar{m})(\bar{\lambda}\bar{\delta}'\bar{v}')/\bar{\lambda}_c < 0. \end{aligned}$$

Because Ω is the product of the system's characteristic roots, it must be the case that there are either three roots with negative real part or one negative root. (Recall that complex roots must occur as conjugate pairs.) Since we already know there cannot be three roots with negative real part, there must be exactly one negative root. This proves that in a neighborhood of long-run equilibrium, our system is saddlepath-stable.

It remains to provide a characterization of the unique stable path. Let θ_1 denote the system's sole negative root, and let $\vec{\omega} = [\omega_{11}, \omega_{21}, \omega_{31}]'$ denote an eigenvector belonging to θ_1. As was shown in the text, c, m, and F must rise or fall together along the saddlepath if the components of $\vec{\omega}$ are all of the same sign. We now prove that this is the case.

First, we note that if one of the components of $\vec{\omega}$ is zero, all must be zero, and so all components must be strictly positive or negative. To see this, suppose $\omega_{11} = 0$. Then, using the last equation of (A11), $-\omega_{11} + r\omega_{31} = \theta_1\omega_{31}$,

and so $\omega_{31} = 0$. But using the second equation of (A11), we find that $\omega_{21} = 0$ as well; and so $\vec{\omega} = 0$. Similar arguments dispose of the possibilities that $\omega_{21} = 0$ and that $\omega_{31} = 0$.

So suppose $\omega_{31} > 0$. From (A11), $-\omega_{11} + r\omega_{31} = \theta_1\omega_{31} < 0$, and so it must be the case that $\omega_{11} > 0$. Again using (A11), we see that $(-\bar{x}_c\bar{m})\omega_{11} + (-\bar{x}_m\bar{m})\omega_{21} = \theta_1\omega_{21}$, implying $\omega_{21} = [(-\bar{x}_c\bar{m})/(\theta_1 + \bar{x}_m\bar{m})]\omega_{11}$. Thus, it must be true that $\omega_{21} > 0$. In this case, therefore, all components of $\vec{\omega}$ have the same sign.

In the case $\omega_{31} < 0$, an identical argument shows that $\omega_{11}, \omega_{21} < 0$ as well. This completes the proof.

Appendix B

This appendix contains a proof that the unique convergent path of the system (19)–(21) is a perfect-foresight equilibrium path. The minor technical issue in the proof is that the state variable a_t may assume negative values, so that the sufficiency theorem for optimal controls (see Arrow and Kurz 1970, p. 49) appears inapplicable. We show that the intertemporal budget constraint (8) implies that a_t must be bounded from below when consumers expect the path of prices implied by the convergent solution to (19)–(21). This fact, in the present context, implies that the sufficiency theorem remains applicable.

Substituting the flow constraint (7) into (8) and integrating by parts, we obtain the inequality

$$\lim_{t\to\infty} a_t e^{-rt} \geq 0. \tag{B1}$$

Condition (B1) may be used to prove the inequality

$$a_t + \int_t^\infty e^{-r(s-t)}\hat{\tau}_s ds \geq 0, \text{ for all } t, \tag{B2}$$

where hatted values are those expected to prevail along the convergent path. Assume (B2) fails at time t'. Integration of (7) implies that for $t > t'$, a_t must satisfy

$$a_t = a_{t'}e^{r(t-t')} + \int_{t'}^t e^{-r(s-t)}[\hat{\tau}_s - c_s - (\hat{\pi}_s + r)m_s]ds, \tag{B3}$$

for any path of consumption and real balances. Multiplying both sides of (B3) by e^{-rt} and taking limits as $t \to \infty$ yields the equality

$$\lim_{t\to\infty} a_t e^{-rt} = e^{-rt'}\left\{a_{t'} + \int_{t'}^\infty e^{-r(s-t')}[\hat{\tau}_s - c_s - (\hat{\pi}_s + r)m_s]ds\right\}. \tag{B4}$$

Because consumption and real balances are nonnegative, the right-hand side of (B4) must be negative if (B2) is false at t'.[17] This contradicts (B1), showing that (B2) must hold at each instant.

Define

$$k \equiv \sup_t \left\{\int_t^\infty e^{-r(s-t)}\hat{\tau}_s ds\right\}.$$

The number k is finite as a result of the boundedness of the sequence $\{\hat{\tau}_s\}$ along the convergent path. From (B2),

$$a_t + k \geq 0, \text{ for all } t. \tag{B5}$$

[17] Our assumptions ensure that $\hat{\pi} + r > 0$ at all times.

It is now straightforward to apply the sufficiency theorem. Redefine the state variable for the household's problem to be $w_t \equiv a_t + k$. The flow constraint (7) becomes $\dot{w}_t = rw_t + \hat{\tau}_t - c_t - (\hat{\pi}_t + r)m_t - rk$. The differential equation system (19)–(21) is unchanged as a result of this modification, and along its convergent path,

$$\lim_{\Delta \to \infty} w_\Delta \lambda_\Delta e^{-\Delta} = 0, \lim_{\Delta \to \infty} \lambda_\Delta e^{-\Delta} \geq 0.$$

Because negative values of the state variable are infeasible, the sufficiency theorem implies that the choices of consumption and real balances implied by the convergent path are indeed optimal choices, given the associated paths of the price level and real transfer payments. The convergent path is therefore a perfect-foresight equilibrium path.

References

Arrow, Kenneth J., and Kurz, Mordecai. *Public Investment, the Rate of Return, and Optimal Fiscal Policy.* Baltimore: Johns Hopkins Univ. Press (for Resources for the Future), 1970.

Brock, William A. "Money and Growth: The Case of Long Run Perfect Foresight." *Internat. Econ. Rev.* 15 (October 1974): 750–77.

Calvo, Guillermo A. "On Models of Money and Perfect Foresight." *Internat. Econ. Rev.* 20 (February 1979): 83–103. (*a*)

———. "Stabilization Rules and the Managed Float: A Search for the Essentials." Discussion Paper no. 23, Columbia Univ., Dept. Econ., 1979. (*b*)

———. "Financial Opening, Crawling Peg and the Real Exchange Rate." Mimeographed. Centro de Estudios Macroeconómicos de Argentina, Buenos Aires, 1980.

Calvo, Guillermo A., and Findlay, Ronald. "On the Optimal Acquisition of Foreign Capital through Investment of Oil Export Revenues." *J. Internat. Econ.* 8 (November 1978): 513–24.

Calvo, Guillermo A., and Rodríguez, Carlos Alfredo. "A Model of Exchange Rate Determination under Currency Substitution and Rational Expectations." *J.P.E.* 85, no. 3 (June 1977): 617–25.

Findlay, Ronald. "An 'Austrian' Model of International Trade and Interest Rate Equalization." *J.P.E.* 86, no. 6 (December 1978): 989–1007.

Fischer, Stanley. "Capital Accumulation on the Transition Path in a Monetary Optimizing Model." *Econometrica* 47 (November 1979): 1433–39.

Flood, Robert P. "An Example of Exchange Rate Overshooting." *Southern Econ. J.* 46 (July 1979): 168–78.

Helpman, Elhanan. "Inflation and Balance of Payments Adjustments with Maximizing Consumers." Seminar Paper no. 129, Inst. Internat. Econ. Studies, Univ. Stockholm, 1979.

Hirsch, Morris W., and Smale, Stephen. *Differential Equations, Dynamical Systems, and Linear Algebra.* New York: Academic Press, 1974.

Koopmans, Tjalling C.; Diamond, Peter A.; and Williamson, Richard E. "Stationary Utility and Time Perspective." *Econometrica* 32 (January/April 1964): 82–100.

Kouri, Pentti J. K. "The Exchange Rate and the Balance of Payments in the Short Run and in the Long Run: A Monetary Approach." *Scandinavian J. Econ.* 78 (May 1976): 280–304.

———. "Balance of Payments and the Foreign Exchange Market: A Dynamic

Partial Equilibrium Model." Cowles Foundation Discussion Paper no. 510, Yale Univ., Dept. Econ., 1978.

———. "Monetary Policy, the Balance of Payments, and the Exchange Rate." In *The Functioning of Floating Exchange Rates: Theory, Evidence, and Policy Implications*, edited by David Bigman and Teizo Taya. Cambridge, Mass.: Ballinger, 1980.

Mundell, Robert A. *Monetary Theory: Inflation, Interest, and Growth in the World Economy*. Pacific Palisades, Calif.: Goodyear, 1971.

Obstfeld, Maurice. "Macroeconomic Policy and World Welfare under Flexible Exchange Rates." Discussion Paper no. 63, Columbia Univ., Dept. Econ., 1980.

———. "Capital Mobility and Devaluation in an Optimizing Model with Rational Expectations." *A.E.R. Papers and Proc.* 71 (May 1981): 217–21.

Razin, Assaf. "Exchange Rates and Price Movements in a World of Capital Mobility." Mimeographed. Foerder Inst., Tel-Aviv Univ., 1979.

Sidrauski, Miguel. "Rational Choice and Patterns of Growth in a Monetary Economy." *A.E.R. Papers and Proc.* 57 (May 1967): 534–44.

Stockman, Alan C. "A Theory of Exchange Rate Determination." *J.P.E.* 88, no. 4 (August 1980): 673–98.

Uzawa, Hirofumi. "Time Preference, the Consumption Function, and Optimum Asset Holdings." In *Value, Capital, and Growth: Papers in Honour of Sir John Hicks*, edited by J. N. Wolfe. Chicago: Aldine, 1968.

Wage and Employment Determination under Trade Unionism: The International Typographical Union

James N. Dertouzos

Rand Corporation

John H. Pencavel

Stanford University and National Bureau of Economic Research

The wages and employment of typographers are examined to see whether they can be usefully characterized as the outcome of a process by which the union maximizes an objective function containing wages and employment and is constrained by a trade-off between these two variables as represented by the employer's labor demand function. Our functional form assumptions permit investigation of some familiar special cases of union behavior. The parameter implications of both the wage bill maximization hypothesis and the rent maximization hypothesis provide inferior explanations of the movement of wages and employment of these workers compared with our more general formulation.

I. Introduction

Although empirical studies measuring the effects of trade unions have multiplied in the last few years, the behavioral underpinnings of

We are indebted to James N. Rosse for sharing with us his deep knowledge of the newspaper industry. Also, we have benefited from comments on a preliminary draft of this paper received from Orley Ashenfelter, Henry Farber, Michael Hurd, Kathie Krumm, Tom MaCurdy, Robert Michael, Daniel Mitchell, Sam Peltzman, Robert Solow, Finis Welch, and an anonymous referee of this *Journal*. Support from the Sloan Foundation to the Department of Economics at Stanford University is gratefully acknowledged.

[*Journal of Political Economy*, 1981, vol. 89, no. 6]
© 1981 by The University of Chicago 0022-3808/81/8906-0005$01.50

unionism remain poorly understood. This is because the issues raised in the literature on trade union goals and wage determination have been mired in fruitless methodological debates or simply have not been pursued at an empirical level. Consequently, although a variety of plausible models of unionism have been proposed, an evaluation of them in terms of their correspondence with observed behavior would rest today on barely more reliable information than was available to Dunlop (1944) when he wrote his seminal work almost 40 years ago. This dearth of research is all the more remarkable when contrasted with the abundant theoretical and empirical work by economists on the behavior of firms and of families. And yet, as Farber (1978) ably demonstrated with his analysis of the United Mine Workers, there is nothing about the trade union as an institution that makes its central features impossible to characterize in a framework analogous to purposive models in economics.[1]

The model of the trade union analyzed in this paper is by no means original; it is usually associated with the work of Fellner (1947) and Cartter (1959), among others. It characterizes the union as maximizing a function in which wage rates and employment are arguments and as being constrained by a trade-off between these two variables represented by the employer's labor demand function. The novel feature of this paper is that we have identified a setting that seems suitable for the operation of the model and have estimated the wage and employment functions that result from this characterization. Therefore, this represents the first empirical application of a model that has been widely discussed in the literature. Moreover, several models of trade union behavior are nested as special cases in our framework so that we are able to provide formal tests of some popular hypotheses. Our purpose is not to champion a specific model but simply to examine whether, in fact, the particular purposive model that is common in the literature provides a viable framework for analyzing union behavior. Only through a careful documentation of a number of empirical studies will the profession be in a position to evaluate the relevance of this model.

In this paper we first describe some basic features of the newspaper industry and of the International Typographical Union (ITU), which is the union with which this paper is concerned. As is well known, this is a powerful, closed-shop union that operates in an industry whose technological features have changed considerably over the last 10–15 years. This paper focuses on a period (from 1946 to 1965) before the drastic changes in the nature of newspaper composition, and during

[1] This opinion differs from the one widely held that, as expressed by Reder (1952), ". . . the behavior of firms and consumers can be easily interpreted as 'maximizing' while that of a union cannot" (p. 34).

these years the model is less likely to have experienced structural change. We provide reasons for believing that bilateral monopoly is not an apt description of the market for typographers, but we cannot claim that our results vindicate this judgment since a bargaining model is not outlined and the estimates are interpreted in terms of the Fellner-Cartter model. Since this model seems to have escaped formal description, we offer such a description in Section III and also introduce our particular functional form assumptions. In Section IV we present our experiences with estimating the model, first with an excellent data set from Cincinnati and then with data on seven other unions. Some conclusions follow in Section V.

II. Institutional Setting

The organization, structure, and traditions of the ITU suggest several features relevant to an analytical model of union behavior. The structure of the ITU parallels that of the printing industry, which mainly consists of geographically segmented product markets. Today there are approximately 1,000 chapters of the union which negotiate individual contracts with local printing establishments, the most important being newspaper firms. In contrast to many industrial unions, the membership of the ITU is remarkably homogeneous. All members undergo the same training, there is no observable prestige differentiation, and education levels and demographic characteristics are quite uniform.[2]

Officers of the ITU on both the national and local levels appear to be extremely responsive to the wishes of their constituent typographers, and the union is frequently described as the most democratic trade union in the United States. Two active political parties compete, and a rather frequent turnover of officers occurs through the electoral process.[3] The union leaders have few incentives to entrench themselves in office. Financial remuneration, even at the highest level, differs little from journeyman wages, and local officers are often workers who perform union duties in their spare time. Negotiated settlements are as likely to affect their own financial status as that of

[2] Lipset, Trow, and Coleman (1956) write that ". . . all of the members of the ITU share a roughly common income and status. . . . The ITU is thus, insofar as this is possible, a community of equals, and there is consequently no 'underprivileged' group in the unions . . . all subskills contribute roughly equal proportions of active leaders and one finds no tendency for a high-status group to monopolize and dominate internal politics" (pp. 142–43).

[3] Indeed, a survey of the largest ITU local, New York no. 6, revealed that over one-half of the membership held some union office at one time or another (see Lipset et al. 1956). Referenda are conducted on all sorts of issues, with voter participation reaching 60 to 70 percent even when the outcomes have marginal impacts on individual members.

the "rank and file." The dichotomy between the interests of officers and of members that assumes importance in some unions appears less relevant to the ITU. Local contracts are constructed and negotiated as prescribed by union laws which have existed for decades and which require membership participation and ratification at each stage.

Over the years, the central organization of the ITU has been concerned with broad issues which have relevance to the union as a whole.[4] In contrast, issues concerning remuneration are negotiated at the local level, and the wage scale negotiated is applied to all journeyman printers. There is evidence that the regular employees are not indifferent to the welfare of the "marginal" or substitute workers.[5] Indeed, as early as 1890, the general laws of the ITU contained provisions which attempted to distribute work evenly to all employees.[6] More recently, the practice has been to fill all vacancies from substitute lists of irregularly employed workers.

Therefore, the ITU is a union with a remarkable democratic tradition, with no clear divergence of interests between different groups within the union, and with an apparent concern both for the employment of its members and for their remuneration. A further attribute deserves mention: Although strikes in the newspaper printing and publishing industry as a whole are not uncommon, those involving the ITU are infrequent.[7] Indeed, for the ITU locals whose data we use below, no strikes occurred during the 1946–65 period.[8] The relative unimportance of strikes is consistent with those models that account for strikes in terms of incomplete information either

[4] These have focused on the union's jurisdiction over printing occupations and the proprietary right of members to employment. Most recently, the ITU has resisted the adoption of new technology which permits electronic and computerized composition of newspaper product except in cases where jurisdiction over new skills and maintenance of employment is permitted (see Kelber and Schlesinger 1967).

[5] For instance, all time accumulated in excess of the amount of hours constituting a regular daily or weekly shift is posted as overtime. When this overtime becomes equivalent to a day's pay, the regular is required to take a day off and engage a substitute as long as one is available. This philosophy of "work sharing" probably reached a zenith during the Great Depression when a national convention adopted a referendum vote approving a law directing all situation holders to put a substitute on work at least 1 day per week.

[6] See International Typographical Union 1964, pp. 312–13. Union membership is sometimes viewed as a legacy to be passed from generation to generation so that family ties among the union members provide a further reason for the typical journeyman to take account of the employment effects of his union's policies (see Lipset et al. 1956, p 206).

[7] For instance, in 1945 (the year prior to the period examined below) only 32 of 1,000 locals engaged in any strikes, and all but one of the struck firms resumed publication within a few days. This information is from the industry's weekly magazine, *Editor and Publisher* (1946, p. 76).

[8] Therefore, it involves not too drastic an assumption if the costs of negotiation and the role of strikes which sometimes figure in models of union behavior (see Mitchell 1972; Atherton 1973) are neglected in this study.

within the union or between the union and the employer: Rates of concession are unknown and the negotiation process reveals information about what to expect from the opposition. However, in the newspaper industry, negotiations have been taking place between the same unions and newspaper firms for many decades, the ITU trade journals regularly publish wage scales for each of the 1,000 locals, and what is sometimes described (Reder 1952, p. 39) as a "mature" collective bargaining relation obtains.

The rarity of strikes reflects in large part the relative vulnerability of the local newspaper firm to a prolonged interruption of production. In particular, the firm's product, the prompt delivery of up-to-date information, cannot be stockpiled to withstand a shutdown so that advertisers and subscribers are apt to turn to alternative media as a result. Once a firm closes down, it is expensive to regain former customers, especially advertisers, who provide an average of about 75 percent of all newspaper revenue and who can turn to ready alternatives (television, weekly newspapers, direct mail, billboards, radio) for their purposes. In contrast, the ITU maintains a defense fund to aid striking local unions, and members have the opportunity to work elsewhere in the event of the newspaper firm being driven out of business. Hence, with the newspaper firm facing a number of immediate competitors in its product market and with the typographers able to draw on a sizable strike fund and to arrange for work in other cities, the setting is not a confrontation between two equally powerful adversaries: The union is clearly in a dominant position vis-à-vis the firm. This provides some support for interpreting the pattern of typographers' wages and employment over time as tracing out the union's preferences.

III. The Conceptual Framework

A General Statement

From a description of the institutional setting, we turn now to a formal statement of the union maximizing model. The union is characterized as behaving as if it possesses a twice continuously differentiable, strictly concave, objective function,

$$U = g\left(\frac{w}{p}, L\right), \tag{1}$$

where w measures the wage rate, p the price level of commodities consumed or the "cost of living," and L union employment. This union is assumed to be involved in producing such a small part of the economy's total output that any effect of its decisions on the overall

price level (p) may be neglected. Hence p is taken to be exogenous to the union, while w and L are endogenous. The union consists of many members so that an issue arises of whose utility function is represented by equation (1). This has been a recurrent theme in the literature on trade union goals, though formally the problem is no different in kind from economic models of family decision making where the family consists of more than one individual or from the theory of the firm where there is more than one owner.[9] The typical ITU local has already been described as one with an unusually homogeneous membership and where conflicts of interest between the union leadership and the rank and file are not pronounced. Therefore, the objective function above is "the" union leader's, and "he" is assumed to integrate the welfare of all the union's members.

The employer's cost-minimizing labor demand function is given by:

$$L = f(w, r_1, r_2, \ldots, r_n, X), \tag{2}$$

where r_i represents the unit price of input i and X is the level of output produced with $n + 1$ inputs by the firm.[10] As is well known, L is a negative function of w, L must be increasing with respect to at least one other input price, L increases with output (provided labor is not an inferior input), and this labor demand function is homogeneous of degree zero in all input prices.

The union is assumed to select w and L to maximize equation (1) subject to the labor demand function given by equation (2). This yields the following first-order condition:

$$-\frac{g_1}{g_2} = p\frac{\partial f}{\partial w}, \tag{3}$$

where $g_1 = \partial U/\partial(w/p) > 0$ and $g_2 = \partial U/\partial L > 0$: The marginal rate of substitution of wages for employment in the union's objective function equals the cost-of-living normalized slope of the employer's labor demand function with respect to the wage rate.[11] This first-order

[9] The problems involved in aggregating over preferences are discussed in Samuelson (1956) and Becker (1974). A rationalization of eq. (1) in terms of a ". . . union as a political institution run by a bureaucracy whose goals are not necessarily identical with those of the membership" appears in Kotowitz and Portes (1974).

[10] Our assumption with the newspaper firms below is that their output is determined primarily by the demographic characteristics of the local media market which are exogenous to a single firm. In fact, we investigate below the possibility that output should be treated as endogenous (see also Rosse 1970 and Dertouzos 1979).

[11] As is well known, this tangency solution between the union's indifference curve and the slope of the firm's labor demand function does not lie on the contract curve (see, e.g., Leontief 1946 or Fellner 1947). This feature may be unrealistic in some employer-union settings, but it is not clear that it is so for typographers. The second-order condition concerns the relative rates at which the union's isoquant and the employer's labor demand function are changing: $p^{-2}g_{11} + 2p^{-1}g_{12}(\partial f/\partial w) + g_2(\partial^2 f/\partial w^2) + g_{22}(\partial f/\partial w)^2 < 0$.

condition together with the employment demand function (2) may be solved to derive the reduced-form equations of the system:

$$w = \phi(p, r_1, r_2, \ldots, r_n, X) \tag{4}$$

and

$$L = \psi(p, r_1, r_2, \ldots, r_n, X), \tag{5}$$

Since an increase in the cost of living reduces the real value of the wage rate without disturbing the employer's structural labor demand function (2), the effect of an increase in p on the wage rate will be opposite in sign to its effect on employment. Other than this, without further restrictions on the form of the union's objective function and the labor demand function, these wage and employment equations are devoid of any unambiguous sign implications.[12] This dearth of qualitative predictions seems to have escaped notice in the extensive literature on this simple maximizing model.

If the reduced-form equations (4) and (5) are substituted back into the union's objective function (1), an expression is derived that relates the *maximum* value (call it V) of the union's objective function to the exogenous variables of the system:

$$V = V(p, r_1, r_2, \ldots, r_n, X). \tag{6}$$

Through the envelope theorem it is straightforward to show that V decreases with respect to p and with respect to the price of any input that is complementary with labor and that V increases with respect to output and with respect to the price of any input that is substitutable for union labor. Moreover, the trade-off in the union's optimizing objective function (6) between increases in the prices of any two inputs is simply represented by the ratio of the derivatives of the employer's labor demand function with respect to each of the input prices:[13] $(\partial V/\partial r_i)/(\partial V/\partial r_j) = (\partial f/\partial r_i)/(\partial f/\partial r_j)$. Since these derivatives of the labor demand function are related to the question of the degree to which this labor is substitutable for (or complementary with) other inputs, this illustrates the relevance of the production technology in constraining the union's opportunities.

[12] The sort of restrictions on eqq. (1) and (2) that generate refutable sign implications may be illustrated as follows. If eq (1) is strongly separable in wages and employment (or, less restrictively, if $g_{12} \geq 0$), if labor is not an inferior factor, and if an increase in output does not reduce the slope of the labor demand function (i.e., if $\partial^2 f/\partial w \partial X \geq 0$), then an increase in output or an increase in the price of a substitute input results in a higher optimal money wage rate.

[13] Also $(\partial V/\partial r_i)/(\partial V/\partial X) = (\partial f/\partial r_i)/(\partial f/\partial X)$.

Particular Functional Forms

In order to implement this framework, at some stage in the analysis some explicit functional forms have to be introduced. These functions should be consistent with the institutional features of the newspaper industry, they should be straightforward to interpret, and they should be computationally manageable. The approach in this paper is to specify particular forms for the union's objective function (1) and for the firm's employment demand function (2) and then to use the first-order condition (3) to solve for the corresponding reduced-form system (4) and (5). Thus the union's objective function defined over real wages and employment is specified as follows:

$$U\left(\frac{w}{p},L\right) = \left(\frac{w}{p} - \gamma\right)^{\theta}(L - \delta)^{1-\theta}. \tag{7}$$

This function, the Stone-Geary, claims an excellent pedigree in empirical work on estimating systems of consumer demand equations, though this represents its first application to union behavior. In fact, it possesses a number of attractive features for our purposes. Though it is not necessary to adhere to this interpretation, following Samuelson (1948), γ and δ may be thought of as "reference" or "minimum" or "necessary" values with the function being homothetic to the point (γ,δ), and $(w/p) - \gamma$ and $L - \delta$ represent "supernumerary" real wages and employment, respectively. The value of θ indicates the relative importance attached to supernumerary wages versus supernumerary employment; only when $\gamma = \delta = 0$ may θ be interpreted as measuring the relative weight of wages and employment in the union's objective function.

One appealing feature of equation (7) in this context is that it nests other models of union behavior as special cases. For instance, Dunlop's model of "maximization of the wage bill for the total membership" (1944, p. 44) is equivalent for the closed-shop ITU to specifying $\theta = .5$ with $\gamma = \delta = 0$ in the objective function.[14] Or the rent maximization hypothesis (Rosen 1970; de Menil 1971) implies $\theta = .5$, $\delta = 0$, and γ equal to the competitive wage. Dynamic elements may be introduced by relating γ and δ to lagged values of real wages and employment so that the union's objectives are those of wage growth and employment growth.[15] Where a union designs its wage policy with

[14] Clearly, with γ and δ equal to zero, a value of θ of 1 yields an objective function that consists simply of the wage rate (a model of union goals sometimes attributed to Simons [1944]), while a value of θ of 0 implies that only employment matters

[15] Cf. "The bargaining process is almost always directed toward an upward or downward change in a pre-existing wage rate and the greatest interest often centers on the magnitude of the change" (Ross 1948, p. 46). Formally, this particular extension of the Stone-Geary was proposed and analyzed by Stone (1966) and Pollak (1970)

reference to what another union has attained or to a nonunion wage, then γ assumes the value of this comparison wage rate.[16] Hence the Stone-Geary function can accommodate a number of alternative hypotheses that have received attention in the literature on union objectives.

The union's opportunity set is defined by the newspaper firm's labor demand function, the expression for which is as follows:

$$L = \alpha_0 + \alpha_1\left(\frac{w}{r_1}\right) + \alpha_2\left(\frac{r_2}{r_1}\right) + \alpha_3 X + \alpha_4 D, \tag{8}$$

where, as before, the r's indicate the prices of nonlabor inputs and X is the firm's output. The negatively sloped labor demand function requires $\alpha_1 < 0$ and, if labor is not an inferior output, then $\alpha_3 > 0$. During our estimating period (1946–65), there was considerable merger activity in the newspaper industry;[17] so, to allow for such mergers causing a shift in the labor demand function, equation (8) includes a dichotomous variable, D, that takes the value of unity when a newspaper merges with another. This parameterization is consistent with a merger resulting in an entirely different cost function for the newspaper, different in its intercept and different in its slope with respect to each of its arguments. This linear-in-the-parameters labor demand function, equation (8), generates tractable estimating equations and yet conforms to what is known about the production technology in the newspaper industry.[18]

The first-order condition for a maximum of (7) subject to equation (8) is $\theta/(1 - \theta) = -[(w - p\gamma)\alpha_1]/[(L - \delta)r_1]$, which may be solved for the reduced-form wage equation expressed in real terms as

$$\frac{w}{p} = \beta_0 + \beta_1\left(\frac{r_1}{p}\right) + \beta_2\left(\frac{r_2}{p}\right) + \beta_3\left(\frac{X \cdot r_1}{p}\right) + \beta_4\left(\frac{D \cdot r_1}{p}\right), \tag{9}$$

where $\beta_0 = (1 - \theta)\gamma$, $\beta_1 = \theta\alpha_1^{-1}(\delta - \alpha_0)$, $\beta_2 = -\theta\alpha_2\alpha_1^{-1}$, $\beta_3 = -\theta\alpha_3\alpha_1^{-1}$, and $\beta_4 = -\theta\alpha_4\alpha_1^{-1}$. Correspondingly, the reduced-form employment equation is

$$L = \beta_5 + \beta_6\left(\frac{p}{r_1}\right) + \beta_7\left(\frac{r_2}{r_1}\right) + \beta_8 X + \beta_9 D, \tag{10}$$

[16] The importance of such comparisons in affecting the goals of trade unions has been stressed by Ross (1948). Again, formally it involves a similar sort of extension of the Stone-Geary function as that analyzed by Pollak (1976).

[17] Almost 8 percent of all cities with daily newspapers had more than one newspaper in 1948. By 1968 this had fallen to less than 3 percent (see Dertouzos 1979). The effect of a newspaper merger on union preferences was also investigated, and the results are reported below.

[18] More information on the newspaper production technology is contained in Rosse (1970, 1977). Also estimated with the Cincinnati data were labor demand functions whose slopes and intercept changed with a merger, but the results did not differ materially from those reported below.

where $\beta_5 = \alpha_0 + \theta(\delta - \alpha_0)$, $\beta_6 = \alpha_1\gamma(1 - \theta)$, $\beta_7 = (1 - \theta)\alpha_2$, $\beta_8 = (1 - \theta)\alpha_3$, and $\beta_9 = (1 - \theta)\alpha_4$. Observe that, since $0 < \theta < 1$, $\alpha_3 > 0$, and $\alpha_1 < 0$, these equations imply a positive correlation between wages and employment, on the one hand, and firm size, on the other hand.[19] These particular functional form assumptions yield reduced-form employment and real wage equations that are homogeneous of degree zero in p, r_1, and r_2. By substituting equations (9) and (10) into the union's objective function (7), we derive an expression that relates the maximum value of this function to the exogenous variables of the system:

$$V(p,r_1,r_2,X,D) =$$

$$-\left(\frac{\theta}{\alpha_1 p}\right)^{\theta}\left(\frac{1-\theta}{r_1}\right)^{1-\theta}[\alpha_1\gamma p + (\alpha_0 - \delta)r_1 + \alpha_2 r_2 + \alpha_3 X r_1 + \alpha_4 D r_1], \quad (11)$$

which is also, of course, homogeneous of degree zero in p, r_1, and r_2.

With the equations of the system described, we turn now to the estimation of the model, first with a data set on the ITU in Cincinnati and then with other union locals.

IV. Empirical Results

Estimating Methods

The structural parameters of our system were estimated by two procedures. The first exploits the fact that the cost of living, p, enters the system through the union's objective function (7) but is excluded from the firm's structural demand equation (8). Thus, using p as an instrumental variable to predict values for (w/r_1), a conventional instrumental variable procedure was applied to equation (8) which provides estimates of the α's. These α's thus estimated were then imposed in estimating the reduced-form wage equation (9) which yields point estimates of the parameters of the union's objective function (namely, θ, γ, and δ). The second estimating procedure involves the simultaneous estimation of the stochastic form of equations (8) and (9) by full information maximum likelihood.[20] Conditional on the specification of the model being correct, this estimator is consistent, asymptotically efficient, and asymptotically normally distributed. In fact, the point estimates derived from applying the two procedures were invariably similar, and consequently only the maximum-likelihood estimates are reported.

[19] For evidence on the association between wages and firm size in the newspaper industry, see Deitouzos (1979).

[20] In particular, we employed the Berndt, Hall, Hall, and Hausman (1974) computational procedure.

The Cincinnati Local

These procedures were first applied to an excellent data set collected for a newspaper, the *Cincinnati Post,* for the years from 1946 to 1965.[21] In estimating the stochastic form of equations (8) and (9), w_t is measured by the hourly wage scale for journeyman printers employed by the *Post* in year t, L_t by the number of full-time typographical workers in the *Post* composing room, p_t by the BLS consumer price index, X_t by the amount of advertising linage (in thousands) sold annually by the *Post*, r_1 by the price of newsprint (in hundreds of dollars per short ton) quoted in national markets, and r_2 by the BLS wholesale price index for machinery and equipment.[22] In 1958, the *Post* merged with one of its two local competitors, the *Cincinnati Times-Star,* and, as a result, a larger and more efficient plant was occupied. The ITU members of the *Times-Star*'s chapel were absorbed into the *Post* composing room so that the employment of typographers rose from 145 averaged over the years 1946–57 to 232 averaged over the years 1948–65. Thus the merger dummy variable, D, in equations (8) and (9) takes the value of zero from 1946 to 1957 and of unity for the later years. The full information maximum-likelihood estimates of the structural parameters of equations (8) and (9) are given in row *a* of table 1.

With respect to the parameters of the union's objective function, the estimate of θ is significantly less than .5 and greater than 0; the union values both employment and wages. The estimate of γ is significantly greater than zero, suggesting that it is not merely the level of wages that matters to the union but, rather, the excess of wages over some reference level. In fact, the estimate of γ of around 2.1 ranges from 26 to 47 percent below the real wage actually received by these workers over these years. In contrast, the point estimate of δ is negative though insignificantly different from zero. According to the estimates in row *a*, the elasticity of the maximum value of the union's objective function with respect to wages (.55) is slightly greater than the elasticity with respect to employment (.45) when evaluated at the sample mean values of the variables. The implied elasticity of substitution of wages for employment at this point is .69. Using a conventional likelihood ratio rest, the null hypothesis that the non-wage arguments of the labor demand function (namely, r_1, r_2, X, and D) are jointly insignificant in the real wage equation (9) may be rejected at the 5 percent level of significance. These results are con-

[21] In the course of performing consulting activities for the *Cincinnati Post,* one of us obtained employment data normally not available through public sources.

[22] In some estimates, the machinery price index was replaced by a measure of the user cost of capital, but the results were essentially unchanged.

TABLE 1

MAXIMUM-LIKELIHOOD ESTIMATES OF THE STRUCTURAL PARAMETERS FOR THE *Cincinnati Post* 1946–65

	θ	γ	γ_0	γ_1	δ	δ_0	δ_1	α_0	α_1	α_2	α_3	α_4	L
a)	.215	2.134	...	...	−129.75	...	...	204.2	−88.72	1.236	3.472	87.86	−57.20
	(.052)	(.716)			(225.36)			(90.6)	(27.30)	(1.489)	(2.467)	(16.16)	
b)	.5*	0*	...	...	0*	...	...	159.1	−75.67	1.018	5.567	60.88	−82.23
								(72.8)	(3.78)	(1.402)	(1.513)	(10.99)	
c)	.5*	...	2.445	−.035	0*	...	...	332.2	−22.66	4.374	2.554	133.22	−64.61
			(.682)	(.470)				(144.4)	(23.40)	(2.350)	(4.708)	(28.33)	
d)	.054	...	1.456	.872	−326.92	...	...	135.4	−70.21	1.610	3.718	77.95	−52.65
	(.079)		(.665)	(.363)	(860.25)			(123.2)	(28.31)	(2.363)	(1.967)	(21.81)	
e)	.490	...	3.965	.491	...	191.31	87.00	201.0	−59.07	−2.268	5.099	77.05	−47.92
	(.206)		(2.196)	(.720)		(49.69)	(14.93)	(71.5)	(39.59)	(1.431)	(2.634)	(18.08)	
f)	.185	...	1.453	.829	...	0*	73.50	147.1	−71.67	1.185	4.855	71.95	−51.27
	(.089)		(.377)	(.323)			(27.27)	(110.6)	(28.17)	(2.213)	(1.965)	(18.02)	

SOURCES — The typographers wage data are taken from selected issues of the *ITU Bulletin* and the advertising linage data from April issues of *Editor and Publisher*. As explained in the text, the employment data were obtained from the research department of Scripps-Howard Newspapers, owners of the *Cincinnati Post*. Published Bureau of Labor Statistics and Department of Commerce publications provided the sources for p, r_1, r_2, and v.

NOTE — L denotes the maximum value of the likelihood function. The mean values of the variables over the estimating period are as follows: $L = 172.3$, $w = 2.880$, $p = .813$, $r_1 = 1.203$, $r_2 = 77.485$, $X = 18.82$, and $v = 1.378$. Numbers in parentheses are estimated asymptotic SEs.

* Indicates constrained parameters.

sistent with the notion that the union takes account of the employment effects of its wage policy.

As for the parameters of the labor demand function, the implied wage elasticity of employment ranges from about -1.8 to -1.0 with an elasticity of -1.23 at the mean values of the variables.[23] The elasticity of employment with respect to output is estimated at .38 at the mean values of the variables.[24] The estimates imply that newsprint and machinery are each substitutes for labor for all observed values of the variables. At sample mean values, the elasticity of the demand for labor with respect to the price of newsprint is estimated to be .77 and that with respect to the price of machinery to be .46.[25] The estimated coefficient on the merger dummy ($\hat{\alpha}_4 = 87.9$) is virtually identical to the difference between the average number of typographers employed after the merger and before the merger (namely, 87).[26]

At this point, our estimates were contrasted with those corresponding to two popular hypotheses of the objectives of unions. First, under the wage bill maximization hypothesis, $\theta = .5$ and $\gamma = \delta = 0$, and the estimates of the labor demand function fitted under these constraints are given in row *b* of table 1. A conventional likelihood ratio test finds that the unconstrained estimates in row *a* represent a significant improvement (at the 5 percent level) over those fitted under the wage bill maximization hypothesis in row *b*. Second, a test of the rent maximization hypothesis requires an interpretation of γ as the wage rate that typographers would earn if they were not or-

[23] It is not straightforward to compare these estimates of labor demand elasticities with others in the literature since these relate to a small group of workers in a single firm, while previous estimates have been at a much higher level of aggregation.

[24] The assumption that output is exogenous in these equations was considered. For instance, errors in measuring output may arise because our index of newspaper output, namely, advertising linage, ignores dimensions of output such as quality and the space devoted to news. For this reason, some instrumental variable estimates were derived where the instruments for newspaper output consisted of the city-zone household population per year, aggregate U.S. advertising expenditures per household in real dollars, and the share of aggregate advertising expenditures earned by television. Clearly, these variables are exogenous to the individual newspaper firm. The resulting estimates of the parameters of the labor demand function were very similar to those in row *a* of table 1.

[25] This finding that newsprint and typographers are substitute inputs is consistent with the stylized facts in the newspaper industry. For as the relative price of newsprint rose, firms found it optimal to conserve on newsprint by producing zoned editions of newspapers, and typographers were required to compose advertising copy for each edition's separate sections.

[26] Augmenting the labor demand function with a linear time trend variable produces similar results to those reported in row *a* (table 1), while the estimated coefficient on the trend variable is about one-quarter its estimated standard error. To be specific, the maximum-likelihood estimates of this system are as follows (with estimated SEs in parentheses): $\theta = .201$ (.053), $\gamma = 2.126$ (.880), $\delta = -130.02$ (283.58), $\alpha_0 = 196.4$ (91.8), $\alpha_1 = -80.99$ (27.59), $\alpha_2 = 1.087$ (1.451), $\alpha_3 = 3.315$ (2.653), $\alpha_4 = 85.14$ (15.89), and with the estimated coefficient on the time trend variable .290 (1.302).

ganized in a trade union. This alternative, nonunion, wage rate is not observed, of course, but a plausible specification is that this alternative wage is a linear function of what some existing nonunion workers receive. In particular, suppose that, in the absence of the ITU, these workers would earn a linear function of the real average hourly earnings of nonsupervisory workers in the retail trade: $\gamma = \gamma_0 + \gamma_1(v/p)$.[27] With this specification for γ, the rent maximization hypothesis implies the restrictions $\theta = .5$ and $\delta = 0$. The estimates corresponding to this formulation are presented in row *c* of table 1 with the unconstrained estimates in row *d*. A comparison of the values of the likelihood function in rows *c* and *d* results in a rejection (at the 5 percent level) of the null hypothesis that the unconstrained estimates are no different from those implied by the rent maximization hypothesis. Therefore, the objectives of the ITU at the *Cincinnati Post* are not adequately described by the maximization of the wage bill or by the maximization of rents.

The estimates in row *d* of table 1 correspond to a specification in which γ is a linear function of an alternative real wage rate, while δ is a constant. A natural generalization is to permit δ to vary, and, on investigation, it was found that δ appeared to take on a higher value after the 1958 merger which resulted in the ITU members of the *Times-Star* chapel being absorbed into the *Post*'s composing room. The final two rows of table 1 augment the specification of $\gamma = \gamma_0 + \gamma_1(v/p)$ with the specification of $\delta = \delta_0 + \delta_1 D$ in row *e*. This more general formulation yields the highest point estimate of θ (namely, 490) that we encountered for Cincinnati. However, it is not clear what importance should be attached to this because the implied estimates of γ and δ are greater than the corresponding mean values of real wages and employment, respectively, so the objective function is not defined at these values. This particular outcome probably results from too restrictive a specification of the movements in γ and δ, but, unfortunately, we are handicapped in our investigation of this matter by insufficient degrees of freedom provided by this data set. In general, of all the coefficients estimated, we found the point estimates of the employment parameter δ to be the most sensitive to slight alterations in the specification of the model.

Estimates that are more consonant with the Stone-Geary specifica-

[27] To forestall any confusion, the assumption here is not that, if they were not unionized, these typographers would be workers in retail trade. The assumption is that if the typographers were not unionized and remained typographers, their wage over time would be some proportion of what nonunion workers in retail trade receive. The v series used relates to the United States as a whole. An hourly earnings series for nonsupervisory workers in the retail trade in Ohio is available only after 1952 and in Cincinnati after 1957. However, for the years from 1953 to 1966, the Ohio series and the U.S. series move very closely together.

tion are given in row f of table 1, but they are obtained only after arbitrarily restricting the parameter δ_0 to a value of zero. According to the results in row f, the elasticity of substitution between real wages and employment in the union's objective function is .305 at sample mean values of the variables. The elasticity of the maximum value of the union's objective function with respect to wages (.97) is virtually identical to the elasticity with respect to employment (.98). The estimate of γ_1 is consistent with the hypothesis that the ITU's reference wage, γ, rises dollar for dollar with increases in the wages of nonunion workers in retail trade. The merger raises δ by almost exactly the amount by which the merger shifts the labor demand function (as given by α_4). At sample mean values, the elasticity of employment with respect to wages is -1.00, that with respect to output is .53, that with respect to the price of newsprint is .55, and that with respect to the price of machinery is .44.

Finally, to illustrate the use of the union's indirect objective function equation (11), we formed the index V_t/V_0, where V_0 and V_t indicate the values of the indirect objective function in the year 1946 and in each subsequent year t, respectively. This ratio, V_t/V_0, may be interpreted as an index number of the union's objective function given the observed values of p, r_1, r_2, and X in year t relative to its value in the base year 1946. Calculating this welfare index with the parameter estimates from row f of table 1 yields the following results: for 1946, $V_t/V_0 = 1.00$; for 1950, 1.46; for 1954, 1.50; for 1958, 1.57; for 1962, 1.55; and for 1965, 1.76. This registers a 76 percent increase in the union's welfare over this period. In contrast, the real wages of these typographers rose by 41 percent between 1946 and 1965 while employment grew 118 percent.

Other Union Locals

To examine the generality of this framework of union behavior, data for the 1946–65 period were collected for seven other local chapters of the ITU. Unlike the data for the *Cincinnati Post*, observations on firm employment are not readily available so that for these seven locals we used data on local union membership instead.[28] The resulting seven union locals range in size from Fond du Lac, Wisconsin,

[28] Data on local union membership may be obtained from the union's publication *Typographical Journal*. These union membership data include those employed in commercial printing establishments as well as those in newspaper firms. We were obliged to use the union membership data; though, to minimize measurement problems, large cities with substantial printing industries (such as New York and Chicago) were excluded from our analysis. For most union locals, commercial printing employment constitutes a very small fraction of the membership.

with a mean journeyman membership of 41 over the 1946–65 period, to the Columbus union with average membership of over 600. In each city but one, the wage and advertising linage variables relate to the single newspaper firm in the city.[29] The stochastic forms of equations (8) and (9) were fitted to the time-series data for each of the seven unions with estimates given in table 2.

As far as the union's objective function parameters are concerned, first observe that the estimates of γ are greater than the mean values of real wages for Elmira and Fond du Lac and the estimates of δ exceed the mean values of employment for Augusta, Elmira, and Fond du Lac. As was the case for the estimates for Cincinnati in row *e* of table 1, this is a consequence (we surmise) of not accounting adequately for movements in the reference levels, γ and δ.[30] For the other four unions, the estimates of γ are significantly greater than zero in all cases, while for Columbus only would the estimate of δ be evaluated as greater than zero at conventional levels of significance. The point estimates of θ range from .15 in Dubuque to .79 in Columbia. Except for Dubuque, the estimates of θ are significantly greater than zero and smaller than unity.[31] Comparing the results for Columbia, Columbus, Dubuque, and Memphis, the tendency is for the estimates of the "reference" parameters, γ and δ, to be greater for the larger unions. These larger unions are located in bigger cities where opportunities for alternative employment (especially in commercial printing) are greater and where wage levels are higher. In particular, the significance of γ suggests that the union's wage policy is determined with reference to some alternative or comparison wage rate, and models that ignore this aspect of union behavior are neglecting what appears to be a consistent finding in this study. In this regard, using a likelihood ratio test, we are able to reject (at even the 1 percent level) for all of the unions the hypothesis that the unconstrained estimates are no different from those implied by the wage bill maximization hypothesis (under which $\theta = .5$ and $\gamma = \delta = 0$).

The estimates of the parameters of the labor demand function are probably best understood in terms of their implications for the elasticity of demand for labor which are given in the final three columns

[29] Columbus is the single exception. In this city the data are aggregated over several newspaper firms. We follow Rosse (1978) in treating agency shop newspapers (i.e., "enterprises with separately owned and produced editorial products but with joint mechanical production and joint sale of advertising and subscriptions" in the same city [p. 3]) as a single firm.

[30] On occasions, permitting the reference levels γ and δ to vary resulted in a likelihood function with no pronounced maximum.

[31] Evaluated at each union's sample mean values, the elasticity of substitution between wages and employment in the union's objective function is estimated at .70 for Columbia, .04 for Columbus, .49 for Dubuque, and .20 for Memphis

requires more research concerns the intertemporal aspect of the model: Clearly, the objectives of unions are defined over the future values of variables as well as their current values, while the employer's immediate response to a negotiated settlement may well differ from his longer-term response. Any serious attempt to model trade union behavior must eventually come to grips with these issues. What has been described in this paper is only an initial step toward that more satisfactory model.

References

Atherton, Wallace N. *Theory of Union Bargaining Goals.* Princeton, N.J.: Princeton Univ. Press, 1973.

Becker, Gary S. "A Theory of Social Interactions." *J.P.E.* 82, no. 6 (November/December 1974): 1063–93.

Berndt, E. K.; Hall, B. H.; Hall, Robert E.; and Hausman, Jerry A. "Estimation and Inference in Nonlinear Structural Models." *Annals Econ. and Soc. Measurement* 3 (October 1974): 653–65.

Cartter, Allan M. *Theory of Wages and Employment.* Homewood, Ill.: Irwin, 1959.

de Menil, George. *Bargaining: Monopoly Power versus Union Power.* Cambridge, Mass.: MIT Press, 1971.

Dertouzos, James N. "Union Objectives, Wage Determination, and the International Typographical Union." Ph.D. dissertation, Stanford Univ., 1979.

Dunlop, John T. *Wage Determination under Trade Unions.* New York: Macmillan, 1944.

Editor and Publisher (February 23, 1946), p. 76.

Farber, Henry S. "Individual Preferences and Union Wage Determination The Case of the United Mine Workers." *J.P.E.* 86, no. 5 (October 1978) 923–42.

Fellner, William J. "Prices and Wages under Bilateral Monopoly." *Q.J.E.* 61 (August 1947): 503–32.

Hartman, Paul T. *Collective Bargaining and Productivity: The Longshore Mechanization Agreement.* Berkeley: Univ. California Press, 1969.

International Typographical Union. *A Study of the History of the International Typographical Union: 1852–1963.* Vol. 1. Colorado Springs, Col.: ITU 1964.

Kelber, Harry, and Schlesinger, Carl. *Union Printers and Controlled Automation.* New York: Free Press, 1967.

Kotowitz, Yehuda, and Portes, Richard. "The 'Tax on Wage Increases': A Theoretical Analysis." *J. Public Econ.* 3 (May 1974): 113–32.

Leontief, Wassily W. "The Pure Theory of the Guaranteed Annual Wage Contract." *J.P.E.* 54, no. 1 (February 1946): 76–79.

Lipset, Seymour M.; Trow, Martin A.; and Coleman, James S. *Union Democracy: The Internal Politics of the International Typographical Union.* Glencoe, Ill.: Free Press, 1956.

Mitchell, Daniel J. B. "Union Wage Policies: The Ross-Dunlop Debate Reopened." *Industrial Relations* 11 (February 1972): 46–61.

Pollak, Robert A. "Habit Formation and Dynamic Demand Functions." *J.P.E.* 78, no. 4, pt. 1 (July/August 1970): 745–63.

———. "Interdependent Preferences." *A.E.R.* 66 (June 1976): 309–20.

Reder, Melvin W. "The Theory of Union Wage Policy." *Rev. Econ. and Statis.* 34 (February 1952): 34–45.

Rosen, Sherwin. "Unionism and the Occupational Wage Structure in the United States." *Internat. Econ. Rev.* 11 (June 1970): 269–86.

Ross, Arthur M. *Trade Union Wage Policy.* Berkeley: Univ. California Press, 1948.

Rosse, James N. "Estimating Cost Function Parameters without Using Cost Data: Illustrated Methodology." *Econometrica* 38 (March 1970): 256–75.

———. "The Daily Newspaper Firm: A 24 Equation Reduced Form Model." Studies in Industry Economics, no. 76., Stanford Univ., Dept. Econ., January 1977.

———. "The Evolution of One Newspaper Cities." Studies in Industry Economics, no. 95, Stanford Univ., Dept. Econ., December 1978.

Samuelson, Paul A. "Some Implications of 'Linearity.' " *Rev. Econ. Studies* 15, no. 2 (1948): 88–90.

———. "Social Indifference Curves." *Q.J.E.* 70 (February 1956): 1–22.

Simons, Henry C. "Some Reflections on Syndicalism." *J.P.E.* 52, no. 1 (March 1944): 1–25.

Stone, Richard. "The Changing Pattern of Consumption." In *Mathematics in the Social Sciences and Other Essays.* Cambridge, Mass.: MIT Press, 1966.

Cross Hedging

Ronald W. Anderson

Columbia University

Jean-Pierre Danthine

Columbia University and University of Lausanne

The paper provides a theoretical description of hedging in futures markets that accounts for the behavior of a broad class of agents. Specific optimal decision rules are derived for agents concerned with the mean and variance of profit. These rules are used to evaluate how optimal cash and futures positions are related to price expectations, the production possibilities, and the number of futures markets available.

I. Introduction

Futures markets in commodities, foreign exchange, and financial assets are thriving with increased participation by a broad spectrum of hedgers and speculators. Despite this fact, there is no satisfactory systematic account of the optimal use of futures markets by hedgers and of the way real output decisions are affected by the availability of futures markets. Conventional descriptions (such as that of Hieronymus [1971]) seem to propose a view of hedging as taking a position in a futures market that is equal in size and opposite to a predetermined position in the cash market. Such a simple view does

Financial support from the Center for Food Studies at Columbia University Graduate School of Business is gratefully acknowledged. Danthine also thanks the members of the Laboratoire de Recherches, Faculté des Sciences de l'Administration, Université Laval, Quebec, Canada, for their hospitality during the summer of 1978 when the research was initiated. We would like to thank an anonymous referee of this *Journal* for helpful comments.

Journal of Political Economy, 1981, vol 89, no 6]
© 1981 by The University of Chicago 0022-3808/81/8906-0001$01 50

not square well with the observed diversity and sophistication of futures markets participants. This paper is an attempt to provide a theoretical description of hedging that could better account for the behavior of a broad class of hedgers. The following facts and results are underlined: (1) Optimal cash and futures positions are normally determined simultaneously. The availability of hedging opportunities affects production or trading decisions. (2) The optimal hedge can be viewed as the sum of a risk-minimizing position and a speculative position. (3) The traditional "routine hedge" (which prescribes the futures position to be the exact opposite of the cash position) in general is suboptimal in that it cannot be identified as a risk-minimizing position or as an optimal speculative deviation from the minimum-risk position. (4) In the absence of transaction costs and of a "perfect hedge," hedgers will take advantage of the better risk-return performance associated with simultaneous trades in a multiplicity of futures contracts. (5) When a limited number of contracts seems appropriate (as a result of transaction costs or indivisibilities), the question of which contract or combination of contracts gives best hedging results arises. This can be answered by making use of available statistical information. (6) When dealing in a good for which no futures exists, a "cross hedge" may be appropriate; the best cross hedge may be calculated in exactly the same way as a standard hedge. (7) The allocation signal to be used by a hedger for his production decision is a linear combination of his expectations and the relevant futures prices.

More specifically, we derive operational formulae for individuals concerned with the mean and variance of their profit. We acknowledge the fact that cash and futures contracts are rarely perfect substitutes, or that the relation of the cash price to the relevant futures price at the time of delivery may not be predicted with certainty (this is known as basis risk); and we recognize that the rarity of a perfect hedge may justify a portfolio approach to hedging, whereby risk reduction (and speculation) is achieved through dealing in multiple contracts. We derive criteria for choosing the best contracts with which to hedge and evaluate the benefits associated with the existence of hedging opportunities.

Previous literature had *raised* some of the deficiencies of the routine hedge prescription. Thus Working (1962) emphasized that the existence of basis risk prevents the elimination of all risk and was led by his observation of hedging practices to the notions of "anticipatory" and "selective" hedging (where deviations from routine hedge are based on price level expectations). Telser (1955), Johnson (1960), Stein (1961), and Peck (1975) have analytical models of optimal hedging that share some of the characteristics of ours. Johnson and

pectations may substitute for cash price expectations in the output decisions. In the special case where the multiple regression coefficients sum to unity ($\beta\iota = 1$), cash price expectations disappear completely from the decision.

V. Adding Futures Markets

In this section we examine how an agent's decisions can be altered by extending the range of available futures markets. Since we posit no fixed costs in transacting in a market and no costs to formulate a probability distribution for the futures prices, it is clear that extending the list of futures markets will not decrease expected utility.[5] Furthermore, the introduction of new futures will induce a restructuring of the entire portfolio so that the effect on the speculative positions is difficult to predict. Consequently, we concentrate our attention on how the optimal cash position may be affected.

Specifically, we consider whether the cash position will increase or decrease in size when the list of futures is extended. The basic result is that when we move from $m - 1$ futures to m futures, the absolute value of the cash position increases if the full vector of futures prices is not expected to change from period 0 to period 1. This result allows us to argue by induction that increasing the list of futures from m to m^* ($m^* > m$) will increase real production when all these futures present no incentive for speculation. The reason for excluding incentives for speculation is that, when such incentives exist, the addition of futures markets may induce a riskier futures position such that the required risk premium in the cash good first-order condition increases.

To formalize the result, let us partition the full vector of m futures according to $f' = (f_1', f_m)$, where f_1 is an $m - 1$ vector. Similarly, let $p_t^{f'} = (p_{1t}^{f'}, p_{mt}^{f})$, and let the variance/covariance matrix of $(\tilde{p}_1, \tilde{p}_{11}^{f'}, \tilde{p}_{m1}^{f})$ be

the cash price and the price of the hedger's "mutual fund" (see eq. [5]). Then the planning price becomes $\Sigma_{01}\Sigma_{11}^{-1}p_0^f - \tilde{B}g$. This is the current price of the hedger's mutual fund minus the expected generalized basis. This interpretation is theoretically attractive. However, defined in this way, the expected basis cannot be interpreted with expectations on price differences only. Hence it does not retain the practical attractiveness of the usual motion of the basis which is due to observation that price differences are more stable than price levels.

[5] It should be emphasized that the analysis of this section takes the current prices and the distributions of future prices as given. General equilibrium effects of adding futures markets cannot be traced out in this context. In a general equilibrium analysis, Hart (1975) shows that adding more futures markets does not necessarily increase welfare

$$\Sigma = \begin{bmatrix} \sigma_{00} & \Sigma_{01} & \sigma_{0m} \\ \Sigma_{10} & \Sigma_{11} & \Sigma_{1m} \\ \sigma_{m0} & \Sigma_{m1} & \sigma_{mm} \end{bmatrix},$$

where σ_{00}, σ_{0m}, and σ_{mm} are scalars; Σ_{11} is $(m-1) \times (m-1)$; and Σ_{01} and Σ_{m1} are $1 \times (m-1)$. We assume Σ is nonsingular. Then net revenue is

$$\bar{\pi} = \bar{p}_1 y - c(y) - (\bar{p}^f_{11} - p^f_{10})' f_1 - (\bar{p}^f_{m1} - p^f_{m0}) f_m. \tag{10}$$

If we consider maximizing a linear mean variance function of $\bar{\pi}$ with respect to y, f_1, and f_m, the necessary conditions are, respectively:

$$\bar{p}_1 - c'(y) - \chi(y\sigma_{00} - \Sigma_{01} f_1 - \sigma_{0m} f_m) = 0, \tag{11}$$

$$p^f_{10} - \bar{p}^f_{11} - \chi(-y\Sigma_{10} + \Sigma_{11} f_1 + \Sigma_{1m} f_m) = 0, \tag{12}$$

and

$$p^f_{m0} - \bar{p}^f_{m1} - \chi(-y\sigma_{0m} + \Sigma_{m1} f_1 + \sigma_{mm} f_m) = 0. \tag{13}$$

Let y^{**}, f_1^{**}, and f_m^{**} be optimal with a full vector of futures; then they solve (11)–(13). Let y^* and f_1^* be optimal with a restricted list of futures; then they solve (11) and (12) with $f_m = 0$. We will show that if $\bar{p}^f_{11} = p^f_{10}$ and $\bar{p}^f_{m1} = p^f_{m0}$, then $|y^{**}| \geq |y^*|$ with strict inequality when Σ_{1m} or σ_{0m} is non-null.

Under these assumptions, (12) implies $f_1^* = y^* \Sigma_{11}^{-1} \Sigma_{10}$. Inserting this into (11), we see y^* satisfies

$$\bar{p}_1 - c'(y^*) - y^* \chi(\sigma_{00} - \Sigma_{01} \Sigma_{11}^{-1} \Sigma_{10}) = 0. \tag{14}$$

Similarly, using (12) and (13) to solve for f_1^{**} and f_m^{**} and inserting these into (11), we can see that y^{**} must satisfy

$$\bar{p}_1 - c'(y^{**}) - y^* \chi \left[\sigma_{00} - \Sigma_{01} \Sigma_{11}^{-1} \Sigma_{10} - \frac{1}{z\sigma_{mm}} (\sigma_{0m} - \Sigma_{m1} \Sigma_{11}^{-1} \Sigma_{10})^2 \right] = 0, \tag{15}$$

where $z = [1 - (1/\sigma_{mm}) \Sigma_{m1} \Sigma_{11}^{-1} \Sigma_{1m}]$. Now it can be shown that $z > 0$.[6]

[6]
$$\begin{aligned} z &= \left(1 - \frac{1}{\sigma_{mm}} \Sigma_{m1} \Sigma_{11}^{-1} \Sigma_{1m}\right) \\ &= \det \left(\Sigma_{11} - \frac{1}{\sigma_{mm}} \Sigma_{1m} \Sigma_{m1}\right) / \det \Sigma_{11} \\ &= \det \begin{pmatrix} \Sigma_{11} & \Sigma_{1m} \\ \Sigma_{m1} & \sigma_{mm} \end{pmatrix} / \sigma_{mm} \det \Sigma_{11} \end{aligned}$$

Here we have repeatedly made use of the properties of determinants of partitioned matrices (see Dhrymes 1970, pp. 570–71). In the last expression σ_{mm}, Σ_{11}, and the term in parentheses are all minors of Σ, a nonsingular variance/covariance matrix by assumption. Thus the last expression is strictly positive.

Then comparing (14) and (15), we see that if σ_{0m} or Σ_{10} is non-null,

$$\bar{p}_1 - c'(y^*) - y^*\chi\left[\sigma_{00} - \Sigma_{01}\Sigma_{11}^{-1}\Sigma_{10} - \frac{1}{z\sigma_{mm}}(\sigma_{0m} - \Sigma_{m1}\Sigma_{11}^{-1}\Sigma_{10})^2\right] \tag{16}$$

is positive if $y^* > 0$ and negative if $y^* < 0$. Given the properties of $c(\cdot)$, we see $|y^{**}| > |y^*|$. If both σ_{0m} and Σ_m are null, expression (16) reduces to (14), and $y^{**} = y^*$.

As an interesting illustration of the above, consider passing from zero to one futures market. In this case y^{**} satisfies

$$(1 - \beta)\bar{p}_1 + \beta(p_0^f - \bar{B}) - c'(y^{**}) - \chi y^{**}\sigma_{00}(1 - \rho^2) = 0 \tag{17}$$

(here we allow the possibility that $\bar{p}_1^f \neq p_0^f$). Similarly, y^* satisfies

$$\bar{p}_1 - c'(y^*) - \chi y^*\sigma_{00} = 0. \tag{18}$$

Using the properties of $c(\cdot)$, it is indeed easy to see that, provided that $(p_0^f - \bar{B})$ does not differ too widely from $\bar{p}$ and that futures and cash prices are correlated ($\rho \neq 0$), $|y^{**}| > |y^*|$. The introduction of futures trading leads to increased real production. However, if futures and cash prices are sufficiently out of line, the reliance on p_0^f for production decisions may suggest more conservative cash positions. On the other hand, if $\rho = 0$, the futures market is of no help for hedging purposes, and (18) reduces to (17). The original risk of the hedger's real activity is intact (σ_{00}); futures trading is limited to placing speculative side bets.

In this latter case, it is clear that the utility derived from futures trading is entirely associated with the speculative opportunities offered. If, however, we abstract from that aspect of futures trading by assuming $\bar{p}_1^f = p_0^f$, it will be clear that utility increments also result from the hedging opportunities offered by futures markets; the size of the increment depends on the degree of correlation. To see this, substitute (5) into (2) with Σ_{11}, Σ_{01}, p_1^f, and f being scalars and differentiate with respect to y and σ_{01} to get

$$d(EU) = \left[\bar{p}_1 - c' + \frac{\sigma_{01}}{\sigma_{11}}(p_0^f - \bar{p}_1^f) - \chi y\left(\sigma_{00} - \frac{\sigma_{01}^2}{\sigma_{11}}\right)\right]dy$$
$$+ \left[y^2\chi\frac{\sigma_{01}}{\sigma_{11}} + \frac{(p_0^f - \bar{p}_1^f)y}{\sigma_{11}}\right]d\sigma_{01}.$$

Under our assumptions, this reduces to $d(EU)/d\sigma_{01} = \chi y^2(\sigma_{01}/\sigma_{11})$. Expected utility is increasing in $|\sigma_{01}|$.

Finally, we derive a monetary measure of the benefit of being able to trade futures. For simplicity we derive this for a single futures case; similar expressions are possible for the multiple futures case. Let y^* and y^{**} be interpreted as above and suppose that the agent is able to trade in futures only by paying a fixed amount k. Let k^* be the amount that would just leave the agent indifferent to trading futures or not trading futures; k^* can be interpreted as the maximum fixed charge that the agent will be willing to pay for the right to trade futures. Within our framework we find

$$k^* = \bar{p}_1(y^{**} - y^*) - [c(y^{**}) - c(y^*)] + \tfrac{1}{2}\frac{(p_0^f - \bar{p}_1^f)^2}{\chi\sigma_{11}}$$
$$+ (p_0^f - \bar{p}_1^f)y^{**}\frac{\sigma_{01}}{\sigma_{11}} - \frac{\chi}{2}\left[y^{**2}\left(\sigma_{00} - \frac{\sigma_{01}^2}{\sigma_{11}}\right) - y^{*2}\sigma_{00}\right].$$

As an example of how this expression could be used, we might imagine a broker who wishes to assess the service charge a potential hedger would tolerate in order to deal in futures. If it is thought that the hedger is unlikely to perceive the speculative opportunities ($p_0^f \cong \bar{p}_1^f$), the amount the hedger is willing to pay is composed of two parts: $\bar{p}_1(y^{**} - y^*) - [c(y^{**}) - c(y^*)]$ is the expected increment of net revenue on the cash market attributable to dealing in futures, and $\{y^{**2}[\sigma_{00} - (\sigma_{01}^2/\sigma_{11})] - y^{*2}\sigma_{00}\}$ is the increment of risk in the futures regime versus the nonfutures regime. This will tend to be positive if $|\sigma_{01}|$ and c'' are relatively small and negative when $|\sigma_{01}|$ and c'' are relatively large. To sum up, the amount the hedger is willing to pay is his increased expected cash revenue less his increase in risk weighted by his degree of absolute risk aversion.

VI. Restrictions on Stochastic Relations between Cash and Futures Prices

Singular Price Variance/Covariance Matrix

When Σ is singular, knowledge of all futures prices is equivalent, in a certain sense, to knowledge of the cash price. This has interesting implications for both the optimal cash and futures decisions

First, recall that from the properties of partitioned matrices, if Σ_{11} is nonsingular,

$$\det \Sigma = (\sigma_{00} - \Sigma_{01}\Sigma_{11}^{-1}\Sigma_{10}) \det \Sigma_{11}$$
$$= (1 - R^2)\sigma_{00} \det \Sigma_{11}.$$

If Σ is singular, $\det \Sigma = 0$ and we see $\sigma_{00} = \Sigma_{01}\Sigma_{11}^{-1}\Sigma_{10}$ or, equivalently, $R^2 = 1$. Thus when the variance/covariance matrix of prices is singular, the cash price is perfectly correlated with a linear combination of futures prices.

This restriction is closely tied to the notion of the routine hedge. For in this case equation (5) implies $\Sigma_{01}f = (1/\chi)\Sigma_{01}\Sigma_{11}^{-1}(p_0^f - \bar{p}_1^f) + y\sigma_{00}$. The covariance weighted sum of futures positions is decomposed into a speculative component that is proportional to the regression weighted sum of expected futures price changes and a pure hedging component that is proportional to the cash position. The restriction implied by the singularity of Σ stands out more clearly if we dimension (as we are free to do) the cash and futures goods so that $\sigma_{00} = \sigma_{11} = \ldots = \sigma_{nn}$. Then when there is no incentive to speculate we have $\Sigma_{i=1}^{n}\rho_i f_i = y$, where ρ_i is the *simple* correlation coefficient of $\bar{p}_{it}^f$ and $\bar{p}_t$. That is, the correlation coefficient weighted sum of futures must equal the cash position. In this sense, a singular Σ is associated with a "generalized routine hedge." If there is only one futures contract, $n = 1$, the pure hedge is in fact a "proper" routine hedge: $f = y$.

The implication of a singular Σ for the cash position is seen immediately from equation (7) for, making use of (9), it reduces to

$$(1 - \beta\iota)\bar{p}_1 + \beta(p_0^f - \bar{B}_1) - c'(y) = 0. \tag{19}$$

Here the risk premium has disappeared, indicating that the cash decision is no longer affected by the level of risk accepted in the futures portfolio.

No Basis Risk

We have already noted that the interest in the basis is a result of the fact that the difference of prices may be more stable than the price levels. As a limiting case, the cash and futures prices may be random while the basis is not. For example, this would be the case if at the time of the futures market expiration the costs of transportation to the cash market are nonstochastic.

If the vector of bases B_1 is a vector of known constants, $\sigma_{0i} = \sigma_{ii} = \sigma_{00}$ and thus $\rho_i = 1$ for all i. Cash price is perfectly correlated with each futures price. In this case Σ is obviously singular, so that equation (19) holds. We now have the further restriction that the regression coefficients sum to unity. For we have $\sigma_{00} = \Sigma_{01}\Sigma_{11}^{-1}\Sigma_{10} = \beta\Sigma_{10}$ or $1 = \beta\Sigma_{10}(1/\sigma_{00}) = \beta(\rho_1 \ldots \rho_N)' = \beta\iota$. Using this in (19), we see that when there is no basis risk, the optimal cash position is given by $\beta(p_0^f - B_1) - c'(y) = 0$. The planning price is entirely a function of current prices and basis expectations; the expected cash price does not enter the

decision. For the single cross-hedge case, this condition, first derived in Danthine (1978), reads $p_0^f - B_1 - c'(y) = 0$.

VII. Basis Speculation

Working (1962) and others have expressed the view that hedging and speculative behavior are distinguished not by the one being risk avoiding and the other risk taking but by the kinds of risk accepted in the two cases. Speculators take price-level risk, while hedgers take price difference (basis) risk. Thus hedgers are said to speculate on the basis.

Restricting ourselves to the $n = 1$ case, we can give a basis speculation interpretation to the hedger's problem by assuming $c(y) = p_0 y$; that is, the hedger is a storage company with costless storage. When we rewrite equation (1), net revenue becomes

$$\begin{aligned}\tilde{\pi} &= (\tilde{p}_1 - p_0)y - (\tilde{p}_1^f - p_0^f)f \\ &= (B_0 - \tilde{B}_1)y + (p_0^f - \tilde{p}_1)s,\end{aligned}$$

where $s = f - y$ is the excess of the futures position over the routine hedge. The agent speculates on the basis by the amount y and on the price level by the amount s. Working's pure basis speculator arbitrarily sets $s = 0$. Our agent jointly selects y and s, thus, in general, attaining a better risk/return combination. While suboptimal in this sense, pure basis speculation may be defended if the agent possesses information about $\tilde{B}_1$ and var $\tilde{B}_1$ but not the moments of the underlying prices.

VIII. Conclusion

Our descriptions of optimal hedging decisions are specific enough to be implemented and their performance to be evaluated. However, the question of empirical applications raises several issues that we have not addressed explicitly.

First, we have ignored the fact that futures contracts are standardized as to quantity and must be traded in integer multiples. For a small user this discreteness is a serious concern and may eliminate the possibility of multiple cross hedges. Even for the large hedger the discreteness of futures markets implicitly may limit the number of markets that should be considered in the portfolio.

A more important concern is the fact that we have assumed that the agent has knowledge of the relevant moments of the probability distributions of prices. In practice, a considerable effort may be required to estimate these moments, particularly if they are subject to

change over time. The stability of moments (or of some more general structure from which the moments can be derived) is, of course, fundamental to our results, as indeed it is to the notion of hedging. The limited ability to identify such stable relationships is likely to be the most serious constraint in developing very extensive multiple cross hedges. In Anderson and Danthine (1980) we address some of the problems of estimation and illustrate the approach with empirical applications.

While we feel that we have provided an account of hedging in futures markets that is more satisfactory than any other we know, we recognize it is not fully general. One restriction is that we have assumed a nonstochastic production process. Judging from Anderson and Danthine (1978) and Rolfo (1980), which treat production uncertainty, the results of the present paper would not be seriously altered by the relaxation of this assumption. Finally, in assuming only two trading dates, we have suppressed the possibility of changing the futures position many times during the period of holding an open cash position. In Anderson and Danthine (1981) we consider the problem of hedging in a single futures market but with many possible trading dates.

References

Anderson, Ronald W., and Danthine, Jean-Pierre. "Hedger Diversity in Futures Markets: Backwardation and the Coordination of Plans." Working Paper no. 71A, Columbia Univ., Graduate School Bus., 1978.

———. "Hedging and Joint Production: Theory and Illustrations." *J. Finance* 35 (May 1980): 487–98.

——— "The Time Pattern of Hedging and the Volatility of Futures Prices." Working Paper no. 393A, Columbia Univ., Graduate School Bus., 1981.

Danthine, Jean-Pierre. "Information, Futures Prices, and Stabilizing Speculation." *J. Econ. Theory* 17 (February 1978): 79–98.

Dhrymes, Phoebus. *Econometrics*. New York: Harper & Row, 1970.

Hart, Oliver D. "On the Optimality of Equilibrium When the Market Structure Is Incomplete." *J. Econ. Theory* 11 (December 1975): 418–43.

Hieronymus, Thomas A. *Economics of Futures Trading, for Commercial and Personal Profit*. New York: Commodity Res. Bur., 1971.

Johnson, Leland L. "The Theory of Hedging and Speculation in Commodity Futures." *Rev. Econ. Studies* 27 (June 1960): 139–51.

Peck, Anne E. "Hedging and Income Stability: Concepts, Implications, and an Example." *American J. Agricultural Econ.* 57 (August 1975): 410–19.

Rolfo, Jacques. "Optimal Hedging under Price and Quantity Uncertainty: The Case of a Cocoa Producer." *J.P.E.* 88, no. 1 (February 1980): 100–116.

Stein, Jerome L. "The Simultaneous Determination of Spot and Futures Prices." *A.E.R.* 51 (December 1961): 1012–25.

Telser, Lester G. "Safety First and Hedging." *Rev. Econ. Studies* 23, no. 1 (1955): 1–16.

Working, Holbrook. "New Concepts concerning Futures Markets and Prices." *A.E.R.* 52 (June 1962): 431–59.

The Implications of Competition among Jurisdictions: Does Tiebout Need Politics?

Dennis Epple and Allan Zelenitz

Carnegie-Mellon University

The paper investigates whether competition among local jurisdictions is, by itself, sufficient to ensure efficient provision of local public goods. Jurisdictions have fixed boundaries, and each has an entrenched government with the power to tax and supply the public good. Residents can move costlessly among jurisdictions. It is shown that competition among numerous jurisdictions is not sufficient to guarantee public sector efficiency. Though residents can "vote with their feet," land is immobile. Hence, governments can usurp some land rents for their own ends. Increasing the number of jurisdictions limits but cannot completely eliminate the ability to exercise discretionary governmental power.

I. Introduction

Governments are alleged to be too big, inefficiently operated, and unresponsive to the public. The assumption that government activities are undertaken to serve the public has fallen into disrepute.[1] Instead, positive analyses of public activities increasingly take the governmental objective to be that of satisfying a bureaucratic goal such as expenditure maximization or maximization of the utility of the head of the government agency.[2] Such objectives have been assumed to characterize local as well as national governments.[3] Indeed, the idea

We would like to thank the participants in the Political Economy Workshop at the University of Pennsylvania, the Microeconomics Workshop at the University of Virginia, the Public Choice Workshop at Virginia Polytechnic Institute, and an anonymous referee for helpful comments. We thank Robert Inman for his comments and for suggesting the title for the paper.

[1] See, e.g., Downs (1957) and Stigler (1971) for seminal arguments

[2] See Posner 1974; Tullock 1974; and Niskanen 1975

[3] See Wagner and Weber 1975.

Journal of Political Economy, 1981, vol. 89, no. 6]
© 1981 by The University of Chicago. 0022-3808/81/8906-0002$01.50

that local governments may pursue various goals is implicit in analyses of local fiscal variables which assume that the values of these variables are dependent on or influenced by the process of collective choice; the characteristics of the decisive voter(s) as well as the nature and institutional form of local governments are presumed to matter.[4]

At the national level, it is often argued that competition via the electoral process leaves the bureaucrat considerable latitude because voters have limited incentive to learn about candidates or to vote their preferences. At the local level, however, it has been argued that competition among jurisdictions restricts a bureaucrat's ability to pursue policies which do not reflect the residents' desires. In fact, Tiebout (1956) argued that jurisdictional competition would guarantee the same efficiency in the provision of local public goods that competition among firms assures in private goods markets. Tiebout's argument thus appears to imply that the ballot box is unnecessary, that efficient provision of local public services will arise because residents can vote with their feet. If, in fact, jurisdictional competition leads to efficient local goods provision, then analyses of the political choice process would appear to be applicable primarily in those cases in which the relevant set of jurisdictions is small.

In a recent article, Courant, Gramlich, and Rubinfeld (1979) study the choice of fiscal variables by a local government that has been captured by its bureaucracy. Their analysis, which focuses upon decision making in a single jurisdiction, investigates the extent to which the ability of residents to leave the jurisdiction limits the power of its bureaucracy. Their paper provides an interesting analysis of the trade-off faced by a public employees' union in a single jurisdiction whose residents are mobile.

Our analysis focuses on the equilibrium among jurisdictions. We seek to determine how changing the number of jurisdictions (i.e., changing the degree of competition among jurisdictions) constrains tax and expenditure policies of individual governments. We posit alternative assumptions about governmental objectives to examine the consequences of jurisdictional competition.[5]

In particular, we question whether competition among a large number of jurisdictions forces governments to implement policies that are in the best interests of their residents. To answer this question, we deliberately consider a rather extreme version of government

[4] See Barr and Davis 1966; Borcherding and Deacon 1972, Bergstrom and Goodman 1973; Denzau and Mackay 1976; Inman 1978; and Romer and Rosenthal 1978.

[5] Courant and Rubinfeld (1978) explore benefit measurement issues in a model with equilibrium among jurisdictions. However, neither jurisdictional competition nor an investigation of implications of alternative governmental objectives is the focus of their paper.

entrenchment. The government (the mayor, the bureaucracy, or however defined) in each jurisdiction is assumed to choose a tax rate and level of government spending to maximize government profit. The government is limited in its choice of tax instruments by a constitution over which it has no control, but there is no political mechanism by which residents can affect the tax rate or the level of government spending. Residents can, however, move freely among jurisdictions. *We do not claim that local governments actually maximize profits; we do not wish to argue that local governments are entrenched and unresponsive to their electorates. Resident voters, via intrajurisdictional processes, may indeed restrain governmental behavior to be in accord with their own desires. Without denying this possibility, this paper addresses the question of whether interjurisdictional competition can, in itself, accomplish the same end.* Although increasing the number of jurisdictions reduces each government's ability to levy taxes in excess of expenditures, we demonstrate that competition among jurisdictions, even with very many jurisdictions, is not sufficient to prevent individual governments from pursuing policies which are not in the interests of their residents. We thus demonstrate that the right of citizens to choose among many communities cannot completely eliminate governmental monopoly power; that is, competition among governments is not equivalent to competition among firms. A further implication is that the political choice process can matter.

The above results are established in a model in which every government seeks to maximize governmental profit and in which each jurisdiction is equivalent in size and location. Once this result has been established, it is natural to ask whether governments in different jurisdictions can coexist while pursuing different objectives. This question is answered affirmatively even when the number of jurisdictions is very large. In answering this question, we shed additional light on the "capitalization hypothesis" and further clarify the interpretation of tests of the Tiebout hypothesis. Finally, we explain what would happen if all jurisdictions were not homogeneous in size or locational desirability or both. One of the implications is that the presence of a large number of jurisdictions in one portion of the metropolitan area, say, the suburbs, limits the monopoly power of jurisdictions in that portion of the metropolitan area but has little effect on the monopoly power of jurisdictions in other portions of the metropolitan area, say, the central city.

In the section which follows, the model is introduced and some comparative statics are derived for use in the subsequent sections. In Section III, governments are assumed to maximize profits, and the implications of jurisdictional competition are examined. The consequences of jurisdictional competition when objectives differ among

governments are examined in Section IV. In Section V we investigate the case in which jurisdictions of differing size occupy differing portions of the metropolitan area. A summary and conclusions are presented in Section VI.

II. Equilibrium with Multiple Jurisdictions[6]

In a metropolitan area with homogeneous individuals each of whom can freely choose the jurisdiction in which he wishes to reside, no jurisdiction can unilaterally create a level of well-being for its residents different from that available to residents of the remaining jurisdictions (assuming more than one). If, for instance, one jurisdiction chooses a fiscal package (i.e., tax rate and level of government services) which residents find relatively unattractive, they will respond by moving to other jurisdictions. As a result of this movement, housing prices adjust to compensate for differences in the desirability of fiscal packages. In this section, we will set forth more exactly the implications of this assumption that individuals can choose their residential location. The primary purpose of this section is to derive properties of the equilibrium. These properties are used in the analyses in subsequent sections. The analyses in this and the following section are kept relatively simple by the assumption that all jurisdictions are alike both in terms of their natural physical amenities and in terms of their fiscal variables.

Consider, therefore, a metropolitan area with a fixed total land area, $\bar{L}$, in which there are J jurisdictions.[7] The land area is divided equally among the jurisdictions, so that each has a fixed boundary encompassing land area $L = \bar{L}/J$. Housing is produced by competitive firms in the jurisdiction from land and nonland factors via a constant-returns neoclassical production function. The price of nonland factors is assumed fixed and uniform throughout the metropolitan area. The housing-supply function in the jth jurisdiction can then be written as:[8]

$$H^j_s = Lh^j_s(P^j_H), \tag{1}$$

[6] In this and subsequent sections of the paper, we assume an equilibrium exists and we investigate the properties of the equilibrium.

[7] In Sec. V we show that our results are not changed qualitatively if it is assumed that additional land is available at the urban fringe

[8] The housing-supply function is obtained by assuming that housing producers choose the ratio of nonland to land inputs to maximize profits. The production function is assumed to exhibit constant returns to scale. The ratio of nonland to land inputs will be an increasing function of the price of housing, and it follows that the amount of housing supplied per unit of land will increase as the price of housing increases.

where P_H^j is the price of housing in the jth jurisdiction and h_s^j is the supply function of housing per unit of land.

The jurisdiction supplies G^j units of a local public good; each resident is assumed to receive an equal nonshared amount. The good is produced with constant returns to scale from nonland factors, and its unit cost will be taken as unity. Thus, the total cost of the good is proportional to the number of residents. The jurisdiction finances this good by assessment of a flat rate property tax, t^j, on the market value of housing in the jurisdiction. Determination of the levels of G^j and t^j will be discussed in the following sections.

There are $\bar{N}$ identical individuals in the metropolitan area and N^j in the jth jurisdiction. Each individual has a utility function over government services, housing, and numeraire goods (B), $U(G, h, B)$, and faces the budget constraint, $I = P_H h(1 + t) + B$. Individuals in each jurisdiction choose housing quantity to maximize utility, taking G^j and t^j as given. Let $V[G^j, P_H^j(1 + t^j)]$ denote the indirect utility function that results. That is,

$$V[G^j, P_H^j(1 + t^j)] = \max_{h_d^j} U[G^j, h_d^j, I - P_H^j h_d^j(1 + t^j)]. \tag{2}$$

The demand function for housing by an individual in jurisdiction j, given the decision to live in jurisdiction j, that results from this maximization is:

$$h_d^j = h_d^j[G^j, P_H^j(1 + t^j)]. \tag{3}$$

Note that h_d expresses housing demanded per person, unlike h_s, which expresses housing supplied per unit of land.

Each individual is free to move among jurisdictions, and his locational choice is that which maximizes his utility. In equilibrium, no individual has an incentive to move. Thus, utility must be the same in all jurisdictions:

$$V[G^i, P_H^i(1 + t^i)] = V[G^j, P_H^j(1 + t^j)]. \tag{4}$$

All N^j individuals in each jurisdiction must be housed, so

$$N^j h_d^j = L h_s^j, \tag{5}$$

and all individuals in the metropolitan area must live in some jurisdiction,

$$\bar{N} = \sum_{j=1}^{J} N^j. \tag{6}$$

Conditions (5) and (6) can be combined to yield:

$$\bar{N} = \sum_{j=1}^{J} \frac{L h_s^j}{h_d^j} = L \sum_{j=1}^{J} \frac{h_s^j}{h_d^j}. \tag{7}$$

We will assume that the government of each jurisdiction, whatever its objective, adopts a Cournot-Nash strategy, choosing its G^j and t^j treating all other jurisdictions' choices of G and t as parameters (i.e., no attempt will be made to investigate strategies in which one government attempts to determine how changes in its tax rate or service level will affect the choice of tax rates or service levels in other jurisdictions). In the remainder of this section, we derive the comparative static response of P_H^j and N^j in each jurisdiction when t^k and G^k are varied in any one jurisdiction. For notational convenience, we will denote jurisdiction 1 (i.e., $k = 1$) as the jurisdiction whose tax and spending levels are perturbed.[9]

If jurisdiction 1 differentially changes its tax rate while no other fiscal parameter is altered, individuals would move into or out of jurisdiction 1 and out of or into the remaining $J - 1$ jurisdictions. Changes in housing prices will ensure that equilibrium, as described by conditions (4) and (7), is restored. Consequently, totally differentiating (4) and (7), we obtain the response of prices to a change in t^1:

$$-U_B^1 h_d^1(1 + t^1)dP_H^1 + U_B^j h_d^j(1 + t^j)dP_H^j = U_B^1 h_d^1 P_H^1 dt^1, \; j \neq 1.^{10} \quad (8a)$$

$$L \sum_{j=1}^{J} \frac{\partial(h_s^j/h_d^j)}{\partial P_H^j} dP_H^j = -L \frac{\partial(h_s^1/h_d^1)}{\partial t^1} dt^1. \quad (8b)$$

In order to simplify notation, for each jurisdiction, let η^j represent the elasticity of housing demand with respect to the gross-of-tax price

[9] Technically, this model is not "closed" for two reasons. First, land has a positive price in all jurisdictions and the return on land wealth has not been included as income to consumers. Second, if the government obtains tax revenue in excess of expenditures, the profit should be counted as income to some agent. Formally incorporating these elements requires dealing with wealth effects—which are a nuisance and do not shed any light on the issues of interest to us. The return on land wealth could be incorporated in one of two ways. Title to land could be created with each consumer-resident endowed with shares valued at $1/N$th of the aggregate value of the land in the metropolitan area. This would not change our analysis or our results in any essential respect. Alternatively, the utility function of consumers could be assumed to display a constant marginal utility of the numeraire commodity, i.e., to be of the form $U = W(G,h) + B$. With this utility function, variations in wealth arising either from changes in land value or from profits in the government sector would not affect the equilibrium in the market for housing within or across jurisdictions. The fact that wealth effects can be introduced in these ways without affecting our qualitative results reinforces our assertion that they can be assumed away. While closing the model by one or more of the devices mentioned above may be an appealing formalism, the added complexity adds cumbersome detail without in any way illuminating the issues being addressed.

[10] In obtaining (8a) from (4), we have made use of eq. (2). Note that (4) and (2) together imply that the direct utility function evaluated at the optimum choice of housing must be equal across jurisdictions. Hence we simply use the envelope theorem, that residents' housing demand optimally adjusts to the new gross-of-tax price of housing, $P_H(1 + t)$, in each community.

of housing ($\eta < 0$), let θ^j represent the elasticity of housing supply with respect to the net-of-tax price of housing ($\theta > 0$), and let γ^j represent the elasticity of housing demand with respect to government services. In addition, let δ^j represent the functionally determined density of population.

Since in this section all governments are assumed to behave identically and adopt the same strategy and all governments and residents are identical, we assume that all governments will adopt the same G and t in equilibrium. Thus, when t^1 is perturbed, $dP_H^2 = \ldots = dP_H^J$, and the levels of all endogenous variables are the same in all jurisdictions. Using these conditions and the elasticities and ratio defined above, we may rewrite equations (8) as:[11]

$$\frac{dP_H^1}{dt^1} = \frac{P_H[\eta - (J-1)(\theta - \eta)]}{J(1+t)(\theta - \eta)} < 0, \tag{9a}$$

$$\frac{dP_H^j}{dt^1} = \frac{P_H\theta}{J(1+t)(\theta - \eta)} > 0, \qquad j = 2, \ldots, J. \tag{9b}$$

It is apparent from (9a) that an increase in the number of jurisdictions increases the sensitivity of P_H^1 to changes in t^1, that is, as J increases, dP_H^1/dt^1 becomes more negative, and in the limit as $J \to \infty$ it approaches $-P_H/(1+t)$. This implies that in the limit the derivative of the gross-of-tax price, $d[P_H^1(1+t^1)]/dt^1$, approaches zero, so that any change in t^1 is fully offset by a change in P_H^1. As the number of jurisdictions becomes large, the price in other jurisdictions is unaffected by changes in t^1, as is seen by taking the limit of (9b) as $J \to \infty$.

Under our simplifying assumption of identical governments and residents, the comparative static effects on housing prices caused by a change in G^1 can be derived in the same manner as the effects caused by changes in t^1. These comparative statics are:

$$\frac{dP_H^1}{dG^1} = \frac{(J-1)U_G}{JU_B h_d(1+t)} + \frac{\gamma P_H}{JG(\theta - \eta)}; \tag{10a}$$

$$\frac{dP_H^j}{dG^1} = -\frac{U_G}{JU_B h_d(1+t)} + \frac{\gamma P_H}{JG(\theta - \eta)}. \tag{10b}$$

The first term in (10a) reflects the increase in housing prices arising because jurisdiction 1 becomes more desirable to consumers as G^1 rises, so that more consumers wish to enter. The second term indicates the effect of the complementarity ($\gamma > 0$) or substitutability ($\gamma < 0$) of government services and housing. If they are complementary, an increase in G^1 causes an increase in housing demand *per*

[11] Detailed enumeration of the steps in deriving these results and results in subsequent sections is available on request.

consumer—which in turn causes P_H^1 to be bidden up.[12] The terms in (10b) are interpreted similarly.

As the number of jurisdictions increases, equation (10b) indicates that the effects of changes in G^1 on prices in other jurisdictions go to zero. As $J \to \infty$, equation (10a) shows that any rise in G^1 will be offset by a rise in P_H^1 to keep the consumer-resident's utility constant.

The effects of changes in t^1 and G^1 on the population in jurisdiction 1, N^1, are derived by differentiation of (5) and by use of the results in (9) and (10). Performing the necessary differentiation and substitutions yields:

$$\frac{dN^1}{dt^1} = \frac{-(J-1)N^1\theta}{J(1+t)} < 0, \tag{11a}$$

$$\frac{dN^1}{dG^1} = N^1\left[\frac{-\gamma}{G} + \frac{\gamma}{JG} + \frac{(J-1)U_G(\dot{\theta}-\eta)}{JU_BP_Hh_d(1+t)}\right]. \tag{11b}$$

These results are easily summarized. An increase in the tax rate causes a decline in the jurisdiction's population because the fall in the equilibrium net-of-tax price (see [9a]) causes a reduction in housing supply. By contrast, a rise in government services makes the jurisdiction more attractive. As a consequence, the population of the jurisdiction will rise, except in the case where complementarity of housing and government services causes an increase in per capita housing demand large enough to offset the decline in per capita demand that results from the price increase exhibited in (10a).

It should be emphasized that the comparative static results derived above do not depend on any particular objective function for the jurisdictions. They simply characterize an equilibrium in which identical jurisdictions pursue identical Cournot-Nash strategies. They can thus be used to investigate governmental objectives such as maximization of property values, the utility of the median voter, government expenditure, or "government profit."

III. Profit-maximizing Government

In this section we assume that each jurisdiction is controlled by a monopolist whose objective is to maximize governmental profit, defined here as the excess of tax revenue over government expenditure. We will neither defend nor deny the realism of this governmental objective; this choice of extreme objective is simply a convenient vehicle for determining whether jurisdictional competition can

[12] In their study of incidence, Polinsky and Rubinfeld (1978, p. 257) find that this complementarity or substitutability is one factor determining how a change in government services affects land prices and other variables.

eliminate the power of local governments to adopt policies which are not in the best interests of their residents. In practice, voting by residents might prevent such exercise of monopoly power. However, our purpose is to determine whether competition among jurisdictions is by itself sufficient to prevent the exercise of governmental power. Hence, we deliberately assume that the monopolist can set G and t without electoral approval.

Profit to the government of the representative jurisdiction is:[13]

$$\pi = tP_H(t, G)H_s[P_H(t,G)] - N(t,G)G. \tag{12}$$

Superscripts have been dropped since all jurisdictions are identical.

Differentiating (12) and using (1) and the comparative static properties of the functions $P_H(t,G)$ and $N(t,G)$, we can write the first-order conditions for profit maximization as:

$$\frac{\partial \pi}{\partial t} = L\left\{\frac{P_H h_s}{(1+t)}\left[1 - t\theta + \frac{t\theta(1+\theta)}{J(\theta-\eta)}\right] + \frac{G\delta(J-1)\theta}{J(1+t)}\right\} = 0. \tag{13a}$$

$$\frac{\partial \pi}{\partial G} = L\left\{\left[\frac{U_G(J-1)\delta}{U_B J P_H h_d(1+t)} + \frac{\gamma\delta}{JG(\theta-\eta)}\right]\left[tP_H h_d(1+\theta) - G(\theta-\eta)\right] + (\gamma - 1)\right\} = 0. \tag{13b}$$

Canceling L, taking limits as $J \to \infty$, and rearranging the above, we obtain:

$$tP_H h_d = G + \frac{P_H h_d}{\theta}, \tag{14a}$$

$$\frac{U_G}{U_B}\left[1 + \frac{G\eta}{P_H h_d(1+t)}\right] = 1 \quad \gamma \tag{14b}$$

From (14b) it is apparent that the marginal rate of substitution of government services for numeraire goods departs from the ratio of relative prices (which equals one). However, because of the distortion that arises from the property tax on housing, even a welfare-maximizing government would choose a level of public good supply such that the marginal rate of substitution does not equal the price ratio.

From (14a), it is clear, unlike the Cournot result when the number of firms is increased, that government monopoly profits are positive even when the number of jurisdictions is large. Tax revenue per capita, $tP_H h_d$, exceeds government spending per capita, G, by a posi-

[13] We assume that the government does not consume housing

tive amount, related to the value of housing per capita, $P_H h_d$, and to the housing-supply elasticity, θ. Thus competition among governments encompassed by fixed boundaries is not sufficient to eliminate those governments' monopoly power.

For an intuitive understanding of why jurisdictional competition does not eliminate governmental power, consider the following simplified version of the model. Suppose consumers get no utility from government services. Obviously, then, the bureaucratic monopolist will set $G = 0$. The condition determining the profit-maximizing value of t is then (13a) with $G = 0$. Suppose further that θ and η are constant over the relevant range. Then (13a) can be solved for t to yield:

$$t = \frac{J(\theta - \eta)}{\theta[J(\theta - \eta) - 1 - \theta]}. \tag{15}$$

Taking the limit of (15) as $\eta \to -\infty$, it is seen that:

$$\lim_{\eta \to -\infty} t = \frac{1}{\theta}. \tag{16}$$

Thus, when housing demand is perfectly elastic, the profit-maximizing tax rate is determined solely by the elasticity of housing supply independent of the number of jurisdictions. With perfectly elastic housing demand, the price of housing is exogenous, but bureaucratic monopolists can still profit by sharing in the rents of the landowners whose land is confined to some given jurisdiction. Alternatively, suppose that housing is in perfectly elastic supply. Then, from (15):

$$\lim_{\theta \to \infty} t = \begin{cases} \dfrac{-1}{\eta + 1} & \text{if } J = 1 \\ 0 & \text{if } J > 1. \end{cases} \tag{17}$$

With only one jurisdiction, the bureaucratic monopolist can exploit the housing-demand elasticity. If there is more than one jurisdiction, housing can be expanded indefinitely at constant cost in each jurisdiction, and the individual jurisdictions are unable to exploit the elasticity of the demand curve.

Next, consider the effects of an increase in the number of jurisdictions in (15):

$$\lim_{J \to \infty} t = \frac{1}{\theta}. \tag{18}$$

Comparing (16) and (18), we see that the effect of having very many jurisdictions is similar to the effect of having an infinitely elastic housing demand. Competition among jurisdictions prevents a bureaucratic monopolist from exploiting the demand elasticity but not the supply elasticity.

Finally, differentiating t with respect to J in (15) reveals that

$$\frac{dt}{dJ} = -\frac{(\theta - \eta)(1 + \theta)}{\theta[J(\theta - \eta) - (1 + \theta)]^2} < 0. \tag{19}$$

An increase in the number of jurisdictions does monotonically decrease the tax rate.

While these results are surprising, they are easily explained. The government of each jurisdiction is given the exclusive right to tax the property in that jurisdiction. Since boundaries are fixed, a landowner cannot move his land to another jurisdiction by a redefinition of boundaries. Fixity of boundaries prevents the tax on land from being competed away. Jurisdictions compete for residents, and competition among a large number of jurisdictions prevents exploitation of the elasticity of housing demand. Jurisdictions in our model here do not compete for land; consequently, the existence of a large number of jurisdictions does not prevent exploitation of the elasticity of housing supply.

Although the tax exploits only the elasticity of supply when there are many jurisdictions, one should not conclude that the tax merely redistributes wealth from landowners to bureaucrats. The tax reduces the unit price of housing received by suppliers, thereby reducing the supply of housing relative to the no-tax equilibrium. As a result, the gross-of-tax price of housing must be higher and per capita housing consumption lower than in the no-tax equilibrium. The only exception would be if housing supply were perfectly inelastic. For this case, the tax would not affect housing production or consumption.

Our results can be illustrated geometrically if we consider a simplified version of the model in which G is fixed at zero for all jurisdictions. Each curve in the right-hand portion of figure 1 shows combinations of net-of-tax housing prices and tax rates that generate the same gross-of-tax price and hence the same level of utility. The further to the northeast, the lower is the utility level. Let V^0 represent the highest utility attainable, that level which would occur if $t = t^0 = 0$ in all jurisdictions and competitive market forces determined a net-of-tax price of P_H^0. The housing market is depicted in the left-hand diagram. Recall that the housing-supply function per unit of land, h_s, is independent of the number of jurisdictions and of the tax rate. The housing-demand function per unit of land is a product of the housing demand per person, $h_d[P_H(1 + t), \hat{G}]$, and the population density, δ. Hence each jurisdiction exhibits the same housing price, P_H^0, and the same population density, δ^0. The zero tax equilibrium in the housing market is shown as point Z.

Now consider the situation in which all governments set the same positive tax rate, t^1. We show that this lowers utility and results in a

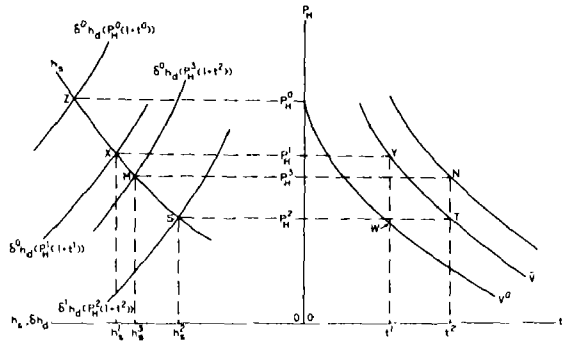


FIG. 1

net-of-tax price lower than P_H^0. If the net-of-tax price were to decline sufficiently to reestablish the original utility level, at point W, the housing-demand function would be unaltered at $\delta^0 h_d[P_H^0(1 + t^0)]$. But with the resulting lower net-of-tax price and an unaltered demand, there would be a housing shortage. With a tax rate of t^1, the housing market could clear only if the net-of-tax price rose above the level at point W. This requires an increase in the gross-of-tax price, hence a decrease in utility below the V^0 level. If P_H did not fall, the housing-demand curve (not shown) would shift downward since $P_H^0(1 + t^1) > P_H^0(1 + t^0)$ and δ is unchanged at δ^0. This would create a housing surplus. Thus, the housing market will clear only if P_H falls below P_H^0, say, to P_H^1. Consequently, a feasible equilibrium with tax rate t^1 is shown as point X in the housing market and as point Y in the utility space. Finally, the tax base per unit of land, with $t = t^1$, is found as the area $OP_H^1Xh_s^1$.

Now, suppose the points X and Y represent the equilibrium when there are many jurisdictions. We illustrate why it would not be in the interest of any of these jurisdictions to unilaterally depart from this equilibrium. We then illustrate why it would be in the interest of a monopoly jurisdiction to choose a higher tax rate.

Suppose a competitive jurisdiction, say jurisdiction 1, were to unilaterally choose a tax rate, say, $t^2 > t^1$.[14] Given residential mobility, the utility level cannot differ across jurisdictions. Thus, the gross-of-tax price in jurisdiction 1, $P_H^2(1 + t^2)$, would equal $P_H^1(1 + t^1)$, which implies $P_H^2 < P_H^1$. Consequently, fewer housing units would be supplied in jurisdiction 1. To restore equilibrium in the housing

[14] We assume in the diagrammatic exposition that competition among jurisdictions does not force all jurisdictions to offer the same tax-service package. This issue is investigated in Sec. IV.

market, it would be necessary for some residents to move to other jurisdictions, lowering the population density in jurisdiction 1, say, to $\delta^1 < \delta^0$. The new hypothetical equilibrium for jurisdiction 1 is illustrated by points S and T. Since it contains but a small fraction of the area's land and population, migration out of jurisdiction 1 would have no effect on the remaining jurisdictions. Thus, the hypothetical equilibrium for the remaining jurisdictions remains at points X and Y.

By increasing its tax rate, jurisdiction 1 decreases its tax base per unit of land to $OP_H^2Sh_s^2$. Should the percentage decline in the tax base be greater than the percentage increase in the tax rate (as suggested by fig. 1), it would not be in the jurisdiction's interest to increase its tax rate. Similarly, were any one jurisdiction to reduce its tax rate below t^2, the resulting increase in the tax base would be more than offset by the decline in the tax rate. Consequently, points X and Y describe the equilibrium for competitive jurisdictions, because there are no incentives for any of the economic agents, governments, residents, or housing suppliers to alter their behavior. Note that, in equilibrium, there are tax revenues collected by each jurisdiction, even when $\bar{G} = 0$; jurisdictional competition is not sufficient to eliminate government profit.

Note that any jurisdiction's incentive to change its tax rate, and hence the equilibrium tax rate, depends on the elasticity of the h_s function. The lower is this elasticity, the greater is the incentive to raise the tax rate. This result can be indicated by considering the case in which h_s is completely inelastic; in particular, assume that h_s rises vertically through point X. If any jurisdiction were to increase its tax rate, any decline in housing demand would not cause a loss of housing units; hence the jurisdiction's tax revenue would increase as long as the percentage increase in the tax rate, $\Delta t/t$, is greater than the percentage decline in the net-of-tax price, $\Delta P_H/P_H$. Since gross-of-tax price is constant, $(P_H - \Delta P_H)(1 + t + \Delta t) = P_H(1 + t)$, and consequently $\Delta P_H/P_H = \Delta t/(1 + t + \Delta t) < \Delta t/t$. A similar incentive to increase t would face each jurisdiction, and t would be continuously forced higher. By altering the supply curve through point X, we can easily show that the incentive to raise t decreases with an increase in the supply elasticity and that a sufficient increase in elasticity creates an incentive to lower the competitive equilibrium tax rate.

Next, suppose that the metropolitan area has but one monopoly jurisdiction. Also assume that the tax rate was initially set at t^1 and that the housing market and utility configurations were initially described by points X and Y. If the monopolist were to choose a higher tax rate, such as t^2, this would increase the tax revenue per unit of land, unlike the situation for the single competitive jurisdiction, because residents cannot escape the monopolist by moving to other jurisdictions. Popu-

lation density throughout the monopolist's jurisdiction would be unchanged at δ^0. An increase in the tax rate to t^2 has two effects. First, assuming momentarily that the net-of-tax price is unchanged at P_H^1, the per capita demand for housing declines, since $P_H^1(1 + t^2) > P_H^1(1 + t^1)$. But with population density fixed at δ^0, this implies an excess supply of housing. Consequently, the second effect is that the net-of-tax price must fall, say to P_H^3. The resulting situation is depicted as points M and N. Note, though, that the decline in the monopolist's tax base per unit of land, to $OP_H^3Mh_x^3$, is substantially smaller than the decline previously indicated for a competitive jurisdiction contemplating the same tax rate increase. Thus, the monopoly jurisdiction would find it profitable to increase its tax rate, even though a competitive jurisdiction would not. Assuming tax revenue changes monotonically with the number of jurisdictions, it follows that tax revenue per unit of land, hence total tax revenue, is inversely related to the number of jurisdictions. With G fixed, government profit is also inversely related to the number of jurisdictions.

IV. Differing Objectives among Governments

As we indicated in the Introduction, data for local jurisdictions have been used extensively for empirical studies of models of government. The implicit presumption of these studies is that competition among jurisdictions does not prevent different governments from pursuing different goals or prevent them from offering different tax and service packages. In this section, we demonstrate that this presumption is correct. In so doing, we provide a theoretical justification for viewing local jurisdictions as natural laboratories for studying models of government. At the same time, we demonstrate that interdependence among local jurisdictions should not be ignored when such studies are conducted. Finally, results in this section make it clear that our result in the preceding section, that competition does not prevent governmental exercise of monopoly power, is not an artifact of our assumption of homogeneous governmental objectives.

As in the preceding section, we assume that an equilibrium exists, and we characterize the properties of that equilibrium as the number of jurisdictions becomes large. It is first necessary to derive the comparative static properties of an equilibrium in which tax and service packages differ across jurisdictions. Conditions (1) through (7) still obtain. The comparative static responses of prices in all jurisdictions to a change in the tax rate in an arbitrarily chosen jurisdiction can be derived from equations (8). There are J equations in (8) from which the J expressions $\partial P_h^j/\partial t^1$ can be derived. The derivation differs from that employed in Section II only in that jurisdictions are not assumed

to adopt identical tax-service packages in equilibrium. By similarly perturbing G in some arbitrarily chosen jurisdiction, say, 1, we can derive the J expressions for $\partial P_H^j/\partial G^1$. Thus, just as in Section II, comparative static responses of price in each jurisdiction to changes in its own fiscal variables and to changes in other jurisdictions' tax rates and service levels can be derived.

By investigating the limiting properties of these comparative statics, we find that, when the number of jurisdictions in a particular subarea becomes very large, the level of utility of residents of any of these jurisdictions is invariant to its choice of tax rate or level of government services. The amount of land in any of these jurisdictions is very small relative to both the total land area and to the land area of jurisdictions having the same amenity level. Consequently, the number of residents migrating into or out of any one of these jurisdictions in response to a change in its tax rate or level of government services is too small to affect the price of housing elsewhere. Hence, utility does not change elsewhere. Rather, the migration induces an adjustment of the housing price within the jurisdiction and that adjustment ensures that the level of utility within the jurisdiction continues to equal the utility level available elsewhere. By contrast, if a jurisdiction has a nonnegligible share of the land in the metropolitan area, migration induced by a change in its fiscal variables is sufficient to change housing prices and utility in all jurisdictions.

While an individual government cannot affect the equilibrium level of utility of its residents when the number of jurisdictions is large, tax rates and service levels are not similarly constrained. Whether changes in these variables are desirable depends upon the government objectives pursued, but it is clear that local governments have latitude in choosing these variables. Put differently, the "market" for local public goods does not force the government to act as a "tax rate taker."[15] This is illustrated in figure 1. Suppose that there are a large number of jurisdictions each of which occupies a small fraction of the land in the metropolitan area. Let $\bar{V}$ be the level of utility realized in equilibrium in all jurisdictions. Consider two of these jurisdictions and assume, for simplicity of illustration, that both offer the same level of government services, $\bar{G}$. The jurisdictions offer differing tax rates, t^1 and t^2, but residents of both jurisdictions enjoy the same level of utility, $\bar{V}$ (shown by the equilibrium points Y and T), and the same

[15] Each of the many jurisdictions faces the constraints as described by $(dP_H^j/dt^1) = (dP_H^j/dG^1) = 0$ and by eqq. (1), (2), and (4). If one, and perhaps only one, of these jurisdictions should be a profit-maximizing jurisdiction, as defined in the previous section, it is relatively easy to show that such a jurisdiction would choose its fiscal variables such that conditions (14) are attained. Hence, choice of t and G by profit maximization is not inconsistent with jurisdictional competition.

gross-of-tax price as all other residents of the metropolitan area. In this equilibrium, jurisdiction 2 has a higher tax rate, a lower net-of-tax price, and a smaller population than jurisdiction 1.

By holding the tax rate constant, jurisdictions offering different levels of government services can be illustrated by a diagram similar to that in figure 1. The diagram would differ only in that G would replace t on the right-hand quadrant, and indifference curves would slope upward. Jurisdictions offering high levels of G would attract more residents than jurisdictions offering low levels of G. As a result, the price of housing would be higher in jurisdictions offering higher levels of government services. Competition will require that utility be the same in all jurisdictions, but equality of utility across jurisdictions does not require equality of tax rates or equality of levels of government services.

The finding that, in equilibrium, utility of residents of a jurisdiction is independent of its government's fiscal decisions is a startling result, one which highlights an important conceptual distinction between residents of a jurisdiction and its landowners. For residents who rent housing, the choices of tax rate and government spending level are matters of complete indifference in equilibrium. Housing prices will adjust to maintain their equilibrium utility level. By contrast, the wealth of landowners is directly affected by the choice of tax rates and level of government service. A resident landowner would clearly support a governmental objective of property value maximization. Such an objective would maximize the wealth of the resident as landowner, and, given his mobility, it would not diminish his utility as a consumer of government services and housing.

The results in this section also have implications for tests of the Tiebout hypothesis. In proposing his test of the Tiebout hypothesis, Oates (1969) argued that differences in tax rates or in the level of public services provided would be capitalized in housing prices. Edel and Sclar (1974) argued precisely the opposite—that in a long-run Tiebout equilibrium, tax rates and government spending levels would be uncorrelated with housing prices. In the model developed here, Oates's interpretation is clearly correct. In equilibrium, differences in tax rates and levels of government services will result in differences in housing prices across jurisdictions. This result does not depend on the existence of a large number of jurisdictions, nor does it depend on the objectives being pursued by individual governments.[16]

[16] It does, however, depend on the assumption that governments do not directly intervene into housing markets (e.g., via zoning) to induce individuals to consume more housing than they would otherwise choose. See Epple (1980) for further discussion of this issue.

The test proposed by Oates may be interpreted as a test of equilibrium condition (4). In equilibrium, with homogeneous residents, the maximum utility level individuals can achieve will be the same everywhere. Denote this utility level $\bar{V}$. Setting $V[G^t, P_H^t(1 + t^t)] = \bar{V}$ and solving for P_H^t, we obtain an equation of the form estimated by Oates:

$$P_H^t = \frac{f(\bar{V}, G^t)}{(1 + t^t)}. \tag{20}$$

In any given equilibrium, $\bar{V}$ is the same across jurisdictions, and hence it is impounded in the constant term in a regression equation. To empirically test this equation, it is necessary that one have observations on at least two jurisdictions. However, the equation holds regardless of the number of jurisdictions. The above result and results in our previous paper (Epple, Zelenitz, and Visscher 1978) indicate the crucial role of assumptions about movability of jurisdictional boundaries. In our earlier paper, we argued that housing prices must necessarily be equal across jurisdictional boundaries when boundaries can be costlessly redrawn. By that argument, the hypothesis of Edel and Sclar is correct, but Oates's hypothesis is correct when boundaries are fixed exogenously.[17]

When the number of jurisdictions is large, changes by any one jurisdiction have no effect on $\bar{V}$. Thus, when the number of jurisdictions is large, differentiation of (20) provides equations that correctly indicate the marginal change in the net-of-tax price of housing which will compensate residents when their jurisdiction changes its tax rate or service level. When the number of jurisdictions is small, a change by any one jurisdiction affects $\bar{V}$. In that case, the appropriate expression for the price derivatives cannot be deduced from equation (20), and, as a result, the compensating alteration in housing price caused by a change in tax rate or service level cannot be deduced from (20).

V. Jurisdictions Differentiated by Size and Location

In preceding sections we considered models in which land at all locations is identical and jurisdictions are identical in size. In this

[17] Note again that the test proposed by Oates is a test of equilibrium condition (4). If individuals are costlessly mobile among jurisdictions, if individuals have full information concerning jurisdictions' alternative fiscal environments, and if jurisdictions do not restrict residents' consumption bundles, then, as Oates clearly saw, equilibrium condition (4) obtains. This is not, however, a test of Tiebout's hypothesis. The latter is a normative statement, concerning the achievement of an amount of the public good, G, which maximizes individuals' utilities. As is clear from the analysis in the paper, there is no assurance that the utility-maximizing level of G is attained for any individual

section we generalize the model to permit jurisdictions to differ in both size and degree of desirability. We argue that our preceding results concerning competition hold when jurisdictions are differentiated by desirability of location. We also discuss how the presence of a large number of small jurisdictions in one part of the metropolitan area affects the monopoly power of a large jurisdiction in the same metropolitan area. Finally, we discuss the effects of permitting new jurisdictions to form at the urban fringe.

Since the method of analysis required here is similar to that employed in previous sections, our discussion is presented without benefit of explicit mathematical argument or proof. Technical details are available on request.

To investigate the effects of differentiation of locations, we assume that the total land area $\bar{L}$ is divided into K subareas or locations indexed by k. Each subarea has amenity (or accessibility) in amount A_k. Variable A is assumed to enter the utility function directly. This specification permits investigation of the effects of varying either the number of jurisdictions at each location or the number of distinct locations.

Equilibrium conditions analogous to (2) through (7) continue to apply. Differences are that amenity level now appears in the utility and demand functions. In addition, supply and demand for housing must be equilibrated at each location. Comparative static results are derived by the same procedure used to obtain equations (9), (10), and (11), and governmental objectives are studied using the approach employed in Sections III and IV.

The implications of increasing the number of jurisdictions offering a given amenity level are found to be similar to those derived in Section III. The presence of a large number of small jurisdictions offering the same amenity level prevents any one of these jurisdictions from exploiting the elasticity of housing demand but not the elasticity of housing supply. The profit-maximizing tax rate for these jurisdictions converges to the inverse of the housing-supply elasticity, a result analogous to that in (18). Tax revenue per unit of land will be higher in locations with a high A_k because the price of housing will be higher at such locations. This is a consequence of the equilibrium condition requiring utility to be constant across jurisdictions.

The presence of a large number of small jurisdictions in one part of the metropolitan area, however, does not prevent a large jurisdiction in another part of the metropolitan area from exploiting both the elasticity of housing demand and supply—a result analogous to that in (15).

To interpret these results, think of two locations in a metropolitan region, one the central city and the other the suburban area. The

above results imply that a large number of jurisdictions in the suburban region are sufficient to prevent individual suburbs from exploiting the elasticity of housing demand. But the relatively large proportion of metropolitan land area occupied by the central city enables its bureaucratic monopolist to exploit the elasticity of housing demand, even if there are a large number of jurisdictions in the suburban region.

Instead of hypothesizing an increase in the number of jurisdictions offering a given amenity level, we might argue that each jurisdiction encompasses land with unique attributes. The implications of increasing the number of distinct jurisdictions occupying a fixed land area are found to be the same as those derived in Section III. The presence of a large number of small jurisdictions, even if each offers a unique amenity, is sufficient to prevent any jurisdiction from exploiting the elasticity of housing demand but not the elasticity of housing supply.

Finally, it might be argued that our results in Section III arise because we assumed a fixed amount of land available for urban use. This conjecture is incorrect. To demonstrate this point, suppose that locations are indexed in order of decreasing desirability, so that jurisdiction K has the lowest amenity level of the existing jurisdictions. Land with amenity level A_{K+1} may then be thought of as the urban fringe, which we assume to be in perfectly elastic supply. With these assumptions, the utility that can be achieved by choosing land with amenity level A_{K+1} is constant at, say, $\bar{U}$.

There are two possible equilibria. All individuals may find the existing K jurisdictions more attractive than the urban fringe. In that case, the urban fringe is uninhabited and all results are unchanged. Alternatively, some individuals may live on the urban fringe. Since all individuals must achieve the same level of utility in equilibrium, the utility level throughout the metropolitan area is $\bar{U}$. However, as demonstrated in Sections III and IV, an entrenched government that must provide its residents an exogenously specified level of utility can still profit by exploiting the elasticity of housing supply. The existence of jurisdictions at the extensive margin does not eliminate land rents in inframarginal jurisdictions, nor does it prevent the taxation of those rents. Therefore, regardless of which outcome occurs, jurisdictional competition, even when residents can relocate on the urban fringe, cannot prevent an entrenched government from exercising monopoly power.

In summary, by the introduction of heterogeneous urban locations, we demonstrate new results and establish that the results of Sections III and IV hold in a much more general setting. Whether the focus be on one urbanized area in isolation or on all urbanized areas within a

nation, whenever rents exist, jurisdictional competition does not prevent an entrenched government from expropriating part of those rents.

VI. Conclusions

The major finding of this paper is that competition among many local jurisdictions is not sufficient to prevent local governments from exercising monopoly power. While we have derived this finding in a model in which government revenues are raised via property taxes, the result also follows if communities are empowered to levy a sales tax (i.e., a tax on the numeraire commodity), an income tax, or a combination of these three taxes.[18]

The feature of our model that is primarily responsible for this result is the assumption of fixed jurisdictional boundaries.[19] A government given taxing powers in a jurisdiction with fixed boundaries can exploit the immobility of land and share in the rents accruing to that land. Mobility of residents across a large number of jurisdictions can prevent individual governments from exploiting the elasticity of housing demand but not the elasticity of housing supply.

These results vindicate researchers seeking to develop a positive political theory of local government behavior. Jurisdictional competition does not predetermine the outcome; Tiebout does need politics. On the other hand, these results highlight the importance of considering the economic as well as the political environment of local governments. As shown in Section IV, competition among a large number of jurisdictions establishes constraints on the choices made by individual governments. Such constraints should not be ignored in modeling decision making by local governments.

References

Barr, James L., and Davis, Otto A. "An Elementary Political and Economic Theory of the Expenditures of Local Governments." *Southern Econ. J.* 33 (October 1966): 149–65.

Bergstrom, Theodore C., and Goodman, Robert P. "Private Demands for Public Goods." *A.E.R.* 63 (June 1973): 280–96.

Borcherding, Thomas E., and Deacon, Robert T. "The Demand for the

[18] See, e.g., Courant's (1977) investigation of the incidence and excess burden of taxes on mobile capital, where the tax rate may be heterogeneous across jurisdictions. He is able to demonstrate that landlords as a class bear at least all of the excess burden of the system of taxes, but he finds that the effects, locality by locality, are not easily characterized.

[19] This is also the feature in Courant et al. (1979) which leads to their conclusions regarding the power of public employees.

Services of Non-Federal Governments." *A.E.R.* 62 (December 1972): 891–901.

Courant, Paul N. "A General Equilibrium Model of Heterogeneous Local Property Taxes." *J. Public Econ.* 8, no. 3 (December 1977): 313–27.

Courant, Paul N.; Gramlich, Edward M.; and Rubinfeld, Daniel L. "Public Employee Market Power and the Level of Government Spending." *A.E.R.* 69, no. 5 (December 1979): 806–17.

Courant, Paul N., and Rubinfeld, Daniel L. "On the Measurement of Benefits in an Urban Context: Some General Equilibrium Issues." *J. Urban Econ.* 5, no. 3 (July 1978): 346–56.

Denzau, Arthur T., and Mackay, Robert J. "Benefit Shares and Majority Voting." *A.E.R.* 66, no. 1 (March 1976): 69–76.

Downs, Anthony. *An Economic Theory of Democracy.* New York: Harper, 1957.

Edel, Matthew, and Sclar, Elliott. "Taxes, Spending, and Property Values: Supply Adjustment in a Tiebout-Oates Model." *J.P.E.* 82, no. 5 (September/October 1974): 941–54.

Epple, Dennis. "What Do Tests of Tax Capitalization Test?" Working Paper, Carnegie-Mellon Univ., Graduate School Indus. Admin., September 1980.

Epple, Dennis; Zelenitz, Allan; and Visscher, Michael. "A Search for Testable Implications of the Tiebout Hypothesis." *J.P.E.* 86, no. 3 (June 1978): 405–25.

Inman, Robert P. "Testing Political Economy's 'As If' Proposition: Is the Median Income Voter Really Decisive?" *Public Choice* 33, no. 4 (1978): 45–65.

Niskanen, William A. "Bureaucrats and Politicians." *J. Law and Econ.* 18, no. 3 (December 1975): 617–43.

Oates, Wallace E. "The Effects of Property Taxes and Local Public Spending on Property Values: An Empirical Study of Tax Capitalization and the Tiebout Hypothesis." *J.P.E.* 77, no. 6 (November/December 1969): 957–71.

Polinsky, A. Mitchell, and Rubinfeld, Daniel L. "The Long-Run Effects of a Residential Property Tax and Local Public Services." *J. Urban Econ.* 5 (1978): 241–62.

Posner, Richard A. "Theories of Economic Regulation." *Bell J. Econ. and Management Sci.* 5, no. 2 (Autumn 1974): 335–58.

Romer, Thomas, and Rosenthal, Howard. "Political Resource Allocation, Controlled Agendas, and the Status Quo." *Public Choice* 33, no. 4 (1978): 27–43.

Stigler, George J. "The Theory of Economic Regulation." *Bell J. Econ. and Management Sci.* 2, no. 1 (Spring 1971): 3–21.

Tiebout, Charles M. "A Pure Theory of Local Expenditures." *J.P.E.* 64, no. 5 (October 1956): 416–24.

Tullock, Gordon R. "Dynamic Hypothesis on Bureaucracy." *Public Choice* 19, no. 2 (Fall 1974): 127–31.

Wagner, Richard E., and Weber, Warren E. "Competition, Monopoly, and the Organization of Government in Metropolitan Areas." *J. Law and Econ.* 18, no. 3 (December 1975): 661–84.

Monetary Expansion and Real Exchange Rate Dynamics

Nissan Liviatan

Hebrew University of Jerusalem

This paper examines the effect of a change in the rate of monetary expansion on the dynamics of the real exchange rate under fully flexible domestic prices and a floating exchange rate. The model is similar in structure to an earlier paper by Calvo and Rodríguez, the main difference being that in our model consumers have an infinite planning horizon and long-run perfect foresight. One of the main findings is that an increase in the rate of monetary expansion has the impact effect of creating a real appreciation of the exchange rate and a deterioration of the balance of payments, contrary to the results of Calvo and Rodríguez.

I. Introduction

This paper analyzes the effects of increasing the rate of expansion of nominal money, μ, on the real exchange rate, e, defined as the ratio of the nominal exchange rate, E, to the price of home goods P_H for an economy operating under a floating exchange rate and fully flexible price level for home goods.

This issue is analyzed in a recent paper by Calvo and Rodríguez (1977) (hereafter CR) in the context of a small two-sector economy (consisting of traded goods and home goods) whose domestic citizens hold both domestic and foreign money, which are the only assets. Their conclusion is that an increase in μ will increase e, that is, cause a real devaluation. Consequently, the impact effect of an increase in monetary expansion is to induce an increase in the production of

I am indebted to Jacob Frenkel for pointing out an error in an earlier version of the paper and for various comments which helped to improve the exposition.

[*Journal of Political Economy*, 1981, vol. 89, no. 6]
© 1981 by The University of Chicago 0022-3808/81/8906-0007$01.50

traded goods and to cause an improvement in the balance of payments.

In the present paper we reexamine the problem within the framework of utility-maximizing behavior and long-run perfect foresight using an approach similar to that of Sidrauski (1967) and Brock (1974). It is shown that a central property of the Calvo-Rodríguez model, namely, the long-run constancy of total assets, no longer holds in the present model. As a result, the impact effect of μ tends to result, under our assumptions, in a real appreciation of the exchange rate, contrary to the CR result. In the final section of the paper the two models are compared in some detail.

II. The CR Model and Results

It will be useful to start with a brief description of the CR model which deals with two commodities—traded goods (T) and home goods (H). The demand functions are

$$c_T = C_T(e,a);\ c_H = C_H(e,a), \tag{1}$$

where $e = (E/P_H)$ is the real exchange rate and a denotes real assets, so that

$$a = m + ef;\ m = \frac{M}{P_H}, \tag{2}$$

where M is nominal domestic money and f is the stock of foreign currency held by domestic citizens. In a full-employment economy output depends only on e, so that

$$q_T = Q_T(e);\ q_H = Q_H(e). \tag{3}$$

Since the home-goods market is always cleared as a result of flexible prices, we have

$$C_H(e,a) = Q_H(e). \tag{4}$$

By using (1) and (3) we can also express the balance of payments

$$\frac{df}{dt} \equiv \dot{f} = Q_T(e) - C_T(e,a) \tag{5}$$

as a function of e and a.

Note that by (2) and (4) we can express a as a function of e, $a = W(e)$, $W' < 0$, so that using (5) we obtain

$$\dot{f} = \phi(e);\ \phi' > 0 \tag{6}$$

by the usual assumptions. The fundamental feature of this relationship, from the point of view of the subsequent analysis, is that it holds

independently of μ. This implies, of course, that the steady-state value of e, derived from $\phi(e) = 0$, is also independent of μ and hence a constant.

The basic relationship of the assets market is

$$\hat{E} = L\left(\frac{m}{ef}\right), \tag{7}$$

where $\hat{E} = (dE/dt)/E$ (we shall use $\hat{X}$ to denote $[dX/dt]/X$), and $L' < 0$. Since $\hat{E} = \hat{e} + \pi$ ($\pi \equiv \hat{P}_H$) and $\hat{m} = \mu - \pi$, we can express (7) as

$$\hat{e} = L\left(\frac{m}{ef}\right) + \hat{m} - \mu. \tag{8}$$

However, since a and e are functionally related, we can express m as a function of e and f. By virtue of (6) and (8) this implies that $\dot{e}$ can be expressed as a function of e, f, and μ, say

$$\dot{e} = \eta(e,f,\mu), \tag{9}$$

where μ is dynamically constant. The assumption of perfect foresight permits identification of "actual" and "expected" rates of change.

Equations (6) and (9) form a dynamic system in e and f with μ as a parameter. As is usual in monetary-growth models with consistent expectations, the dynamic system around the steady state corresponds to a "saddle point" solution. Along the saddle path of the linearized system f follows a linear first-order differential equation, which in turn implies that f converges monotonically to the steady-state solution.

Suppose we start at a steady state and increase μ. Let us assume that this causes an increase in the steady-state value of f (denoted $\bar{f}$). We know, however, that in the course of adjustment to the new equilibrium f must increase monotonically at a diminishing rate, as it should be for a linear differential equation.[1] Hence if $\dot{f}_o$ (the initial value of $[df/dt]$) was zero before the change in μ it must become positive as μ and $\bar{f}$ increase. However, since $\dot{f}$ and e are functionally related by (6), the increase in $\dot{f}_o$ must be associated with an increase in e_o, which is the result proved by CR. In economic terms, in order to support the accumulation of f through the balance of payments we need an initial real devaluation. This argument can be summarized by

$$\text{sign}\,\frac{de_o}{d\mu} = \text{sign}\,\frac{d\dot{f}_o}{d\mu} = \text{sign}\,\frac{d\bar{f}}{d\mu}. \tag{10}$$

[1] The equation for f is one of the form $f(t) = \bar{f} + be^{\lambda t}$, $\lambda < 0$, and $b = f_o - \bar{f}$. Hence $\dot{f} = \lambda be^{\lambda t}$ and $\dot{f}_o = \lambda b$. After the increase in μ we have $\dot{f}_o = \lambda(\bar{f}_o - \bar{f}_1) > 0$ since $\lambda < 0$ and $\bar{f}_1 > \bar{f}_o$.

Since the steady-state value of e is independent of μ and since there is a unique relationship between a and e that is consistent with home-goods equilibrium, it follows that the steady-state value of assets is also independent of the rate of monetary expansion. The higher steady-state rate of depreciation that is associated with the higher μ implies that in the new steady state a higher proportion of the portfolio will be held in the form of foreign exchange so that $\bar{f}$ is raised. Consequently, $\bar{f}_0$ and e_0 increase.

The constancy of the total value of assets across steady states is central to the CR thesis. It may be noted, however, that an increase in μ represents in some sense a reduction in the return on assets in general, which may lead in a different model to reduction in their full equilibrium value. In order to focus attention on this aspect we shall formulate a long-run perfect-foresight model where the relation between the rate of return on assets will have to be balanced against the subjective rate of time preference, as in optimal growth models.

III. The Long-Run Perfect-Foresight (LRPF) Model

We assume that the supply side is the same as in the CR model. The difference is in the treatment of demand for commodities and assets. We wish to base the consumers' behavior on utility analysis. In order to account for differential demand for m and f we shall follow the procedure of including the liquidity services provided by both assets in the utility function (in this respect, we extend the Sidrauski [1967] approach to two assets).

An important assumption that we shall use is that the instantaneous utility function is separable[2] as far as consumption of ordinary commodities and liquidity services is concerned. It is this assumption which will enable us eventually to derive a relationship between $\dot{f}$ and e which is independent of μ, as in (6). The utility integral to be maximized by the representative consumer can then be written as[3]

$$\int_0^\infty e^{-\delta t}[u(c_H,c_T) + V(m,ef)]dt, \tag{11}$$

where u and V are strictly concave functions and $\delta > 0$.

As in CR, there are only two assets—m and f. There are no bonds and hence no borrowing. The only way the consumer can spend in excess of his income is by running down his stocks of m and f ("bor-

[2] The separability in (11) is stronger than necessary. What is required is that the marginal rate of substitution between C_T and C_H in $u(\ )$ be independent of the assets (i.e., separability in the Leontief sense).

[3] $V(\cdot)$ can be expressed more generally as $V(m,f,e)$ without affecting our results

rowing from himself"), so that there is an upper bound on his current consumption possibilities. This assumption is, in fact, required for the existence of a solution.

The income receipts consist of the value of output $q = q_H + eq_T$ and of transfer payments by the government s. We may then express the accumulation of assets ($\dot{a}$) as

$$\dot{a} = q + s - \Pi m - (\Pi - \hat{E})ef - (c_H + ec_T), \tag{12}$$

where Πm and $(\Pi - \hat{E})ef$ represent "depreciation" terms. Clearly, when $(\hat{E} - \Pi)$ is positive we may regard $(\hat{E} - \Pi)ef$, or $\hat{e}(ef)$, as an "income" term. In fact, in the present model $\hat{e}$ often plays a similar role to the "real rate of return on capital" in monetary-growth models. Using the definition $a = m + ef$ we may rewrite (12) as

$$\dot{a} = q + s + \hat{e}a - (\hat{e} + \Pi)m - (c_H + ec_T), \tag{13}$$

where $\hat{e}$ represents the "real interest rate" and $(\hat{e} + \Pi)$ represents the nominal one.

The optimization problem is then to maximize (11) subject to (2), (13), and initial assets (a_o). It should be noted that from the point of view of the maximizing individual e is a market parameter and the two assets are fully interchangeable at any instant. From the Euler conditions we obtain

$$\frac{V_m - V_F}{u_H} = \hat{e} + \Pi, \tag{14}$$

$$\frac{u_T}{u_H} = e, \tag{15}$$

$$\frac{V_F}{u_H} + \hat{e} - \delta = -\frac{\dot{u}_H}{u_H}, \tag{16}$$

where subscripts of u and V denote taking partial derivatives with respect to the indicated variables (e.g., $u_T = [\partial u/\partial c_T]$, $V_F = [\partial V/\partial(ef)]$, etc.).

We may then rewrite (14) as

$$\hat{e} = (-\Pi) + \frac{V_m - V_F}{u_H},$$

which means that along the optimal path the real rate of return on F must equal the real rate of return on m (i.e., $-\Pi$) plus a liquidity premium. Condition (16) is the usual condition of optimal growth models where the role of the "real rate of return on capital" is now taken by $[(V_F/u_H) + \hat{e}]$, which equals the real market rate of return on F plus the nonpecuniary liquidity element.

Let us examine first whether in the present model we may still derive an autonomous relationship between $\dot{f}$ and e. Since in equilibrium $C_H = Q_H(e)$, we can write (15) as

$$\frac{u_T[Q_H(e),c_T]}{u_H[Q_H(e),c_T]} = e, \tag{17}$$

which can be solved to yield $c_T = C_T(e)$. From the balance of payments we then have

$$\dot{f} = Q_T(e) - C_T(e) = \Psi(e), \tag{18}$$

which proves that we still have a functional relationship between f and e which is independent of μ.

The sign of $[dC_T(e)/de]$ is generally ambiguous. If, however, we take the example of a Cobb-Douglas type of utility function, then $(u_T/u_H) = \beta(c_H/c_T)$ (β a constant), and hence by (18) $c_T = [\beta Q_H(e)/e]$ so that $[dC_T(e)/de] < 0$. The same is true for any homothetic $u(\cdot)$ function. In any case, whatever the sign of the latter derivative, we may assume that the sign of Ψ' in (18) is dominated by $Q_T'(e)$, so that $\Psi' > 0$ (clearly if $[dC_T/de] < 0$ this tendency is only strengthened). Thus the properties of $\Psi(e)$ are basically the same as those of $\phi(e)$.

Since in equilibrium c_T can be expressed as a function of e alone and since the same is true of c_H, it follows that u_H and u_T are, in equilibrium, functions of e only, say $u_H(e)$ and $u_T(e)$. This implies that (16) can be expressed as

$$\dot{e} = \theta(e,m,f). \tag{19}$$

By a similar consideration, and using $\Pi = \mu - (\dot{m}/m)$, we can express (14) as $G(\dot{m},m,e,e,f,\mu) = 0$. Substituting (19) in the foregoing relationship and solving for $\dot{m}$ we obtain

$$\dot{m} = \gamma(e,m,f,\mu) \tag{20}$$

Equations (18), (19), and (20) form the dynamic system for e, m, and f for the economy under competitive equilibrium.

When we consider the system at some "initial" point of time, then f_0 is momentarily fixed. However, the initial values e_0 and m_0 are free to be determined by the market since all prices are flexible. We assume that in the LRPF (long-run perfect-foresight) model e_0 and m_0 are determined uniquely so as to set the economy on the "saddle path."

In terms of the linearized system around the steady state, we can express the solution for any of our variables, say for f, in the form

$$f(t) = \bar{f} + b_1e^{\lambda_1 t} + b_2e^{\lambda_2 t} + b_3e^{\lambda_3 t}, \tag{21}$$

where the b's are determined by the parameters of the system and by the initial values, f_0, m_0, and e_0. The fact that e_0 and m_0 are determined

so as to set the system on the stable trajectory leading to the steady state means that two of the three λ's are positive (or have positive real parts in the case of complex roots), and therefore their corresponding b_i's have to be set equal to zero by appropriate values of e_0 and m_0.

Furthermore, the fact that the "saddle path" is stable means that the remaining λ (say λ_1) is negative. Since on the saddle path

$$f(t) = \bar{f} + b_1 e^{\lambda_1 t}, \tag{22}$$

we see that b_1 is determined by $f(0) - \bar{f}$. This leads to the fundamental conclusion that the convergence of $f(t)$ is monotonic, as in the CR model.

We turn now to consider the effect of an increase in μ on the real exchange rate (starting from a steady-state position). Since in a steady state $\dot{f} = 0$, we see immediately from (18) that $\bar{e}$ is uniquely determined, since $\Psi' > 0$. This corresponds, of course, to the slope-of-transformation curve between H and T at the point of tangency with the indifference curves relating to the two commodities. Thus μ has no effect on the long-run real exchange rate.

In order to analyze the impact effect of μ on e_0 we make use again of the fact that $f = \Psi(e)$ is independent of μ, and of the property of monotonic convergence of f. As we saw in connection with the CR model, this implies sign $(de_0/d\mu)$ = sign $(df_0/d\mu)$ = sign $(d\bar{f}/d\mu)$. Let us then determine $(d\bar{f}/d\mu)$, which is the key to the impact effect. The sign of this derivative depends critically on the sign of the mixed derivative V_{mF}. The usual assumption in these cases is that $V_{mF} > 0$; that is, the two assets behave like "cooperant" factors of production in the theory of the firm.

In steady states $\bar{e}$ is a given constant, while conditions (14) and (16) reduce to

$$\frac{V_m - V_F}{u_H} = \mu, \tag{14'}$$

$$\frac{V_F}{u_H} = \delta. \tag{16'}$$

Differentiating this system with respect to m, f, and μ we obtain the following solutions:

$$\frac{dm}{d\mu} = \frac{V_{FF}u_H}{\Delta}; \quad \frac{df}{d\mu} = -\frac{V_{Fm}}{\Delta}\frac{u_H}{\bar{e}} \tag{23}$$

$$\Delta = V_{mm}V_{FF} - V_{FM}^2.$$

Since $V_{FF} < 0$ and $\Delta > 0$ by the concavity of $V(\cdot)$ we find that $(dm/d\mu) < 0$ as expected. We also find that under the "cooperancy" assumption $(df/d\mu) < 0$. Since this refers to $\bar{f}$ of steady states, it follows from

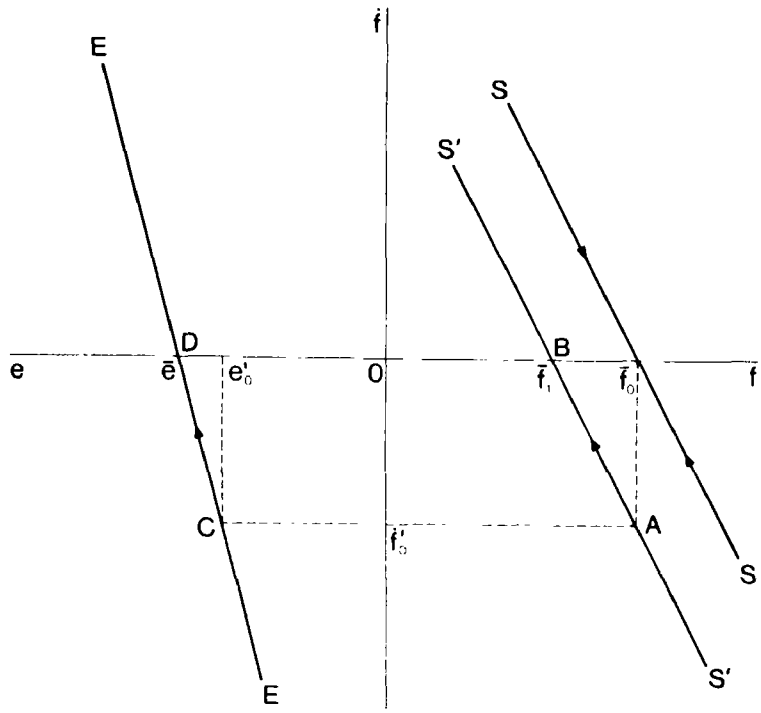


FIG. 1

the earlier remarks that $(df_0/d\mu)$ and $(de_0/d\mu)$ are negative as well. Thus an increase in μ leads to an appreciation of the real exchange rate contrary to the CR result.

These results are illustrated graphically in figure 1, where SS represents the relationship between $\dot{f}$ and f along the saddle path, while EE represents $\dot{f} = \Psi(e)$. Note that e increases from right to left on the horizontal axis. As μ increases we have a new $\dot{f}(\bar{f}_1)$ and a new $S'S'$ curve passing through it. However, since momentarily $f = f_0$, we have $\dot{f}_0$ decreasing from zero to $\dot{f}'_0$. This corresponds on the EE curve to a decrease in e_0 from $\bar{e}$ to e'_0. The dynamic paths are represented by movements along AB and CD. Thus along the dynamic adjustment path we have real depreciation associated with a negative trade balance, but e is still below $\bar{e}$ throughout.

IV. A Comparison of the Alternative Models

The key to understanding the difference in the results obtained in the two models is in the different effect of μ on the steady-state value of f.

The negative effect of μ on $\bar{f}$ in the LRPF model and its implications can be explained as follows.

In steady states the real rate of return on F, that is, $\hat{e}$, is identically zero, while the real rate of return on m is given by $-\mu$. Hence an increase in μ reduces the real rate of return on money and therefore reduces $\bar{m}$. By the cooperancy assumption this reduces the marginal utility of f, and consequently its equilibrium value is reduced. Under long-run perfect foresight the decumulation of f must start immediately so that $\dot{f}_0$ decreases. However, a deterioration of the balance of trade is compatible only with an appreciation of the real exchange rate. Hence $(de_0/d\mu) < 0$.

Since the cooperancy assumption is crucial for our result, we must discuss it in greater detail and interpret it, for the sake of comparison, in the context of the CR model.

Note that (V_m/u_H) and (V_F/u_H) can be interpreted as the subjective rates of return on m and F, which will be denoted by ρ_m and ρ_F, respectively. Condition (14′) can then be written as $\rho_m - \bar{\pi} = \rho_F$ $(\mu = \bar{\pi})$, which indicates that the ρ_F must equal ρ_m after allowing for the inflationary depreciation of real balances. Thus we have essentially a condition of equality of the "net" rates of return properly defined. It can therefore be regarded as a portfolio allocation condition analogous to condition (7) in the CR model.

Indeed, in (14′) we have a relationship between m and F in steady states that must be consistent with μ $(= \hat{E})$, as is the case in (7). If we accept this analogy between (14′) and (7) it seems reasonable to require that for a given μ, m and F should be positively related in (14′). Writing (14′) as

$$\frac{V_m(m,F)}{u_H} - \delta = \mu, \tag{14''}$$

we see immediately that a positive relationship between m and F implies $V_{mF} > 0$; that is, the factors must be cooperant. Thus the cooperancy assumption seems to be in line with CR's equation (7).

While (14′) relates the rates of return to each other, (16′), which can be written as $\rho_F(m,F) = \delta$, pins down the rate of return on F to the given subjective rate of time preference. Thus (16′) helps to determine the "scale" of assets. As μ increases and m decreases, we have to determine what happens to the value of F which is required to maintain $\rho_F = \delta$. In this connection we note that if m and F are positively related in (14″), this must be the case for (16′) as well (since $V_{mF} > 0$ and V is concave).

Note that while (14′) plays a similar role to that of (7) in the CR model, condition (16′) is the one which replaces their assumption of the constancy of $m + F$ across steady states. In the present model the

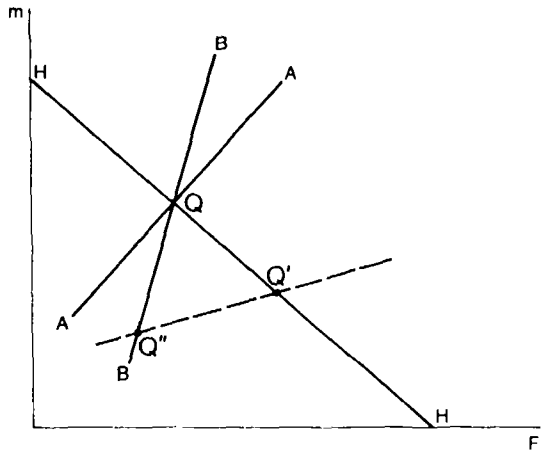


FIG. 2

constancy of the value of total assets is replaced by the condition of constancy of the net rate of return, that is, $\delta = \rho_F = \rho_m - \Pi$, which has quite different implications.

The difference between the models can now be clarified by means of figure 2. Suppose that the AA curve represents the portfolio balance condition given by (7) in the CR model and by (14″) in the LRPF model. To close the model, Calvo and Rodríguez introduce the condition that $m + F$ is constant as represented by the HH curve.

In the LRPF model the system is closed in a different manner, namely, by the condition $\rho_F = \delta$, that is, by (16′) as expressed by BB. The latter curve must be upward sloping since the assets are cooperant, and it is steeper than AA as a result of the concavity of $V(\cdot)$. The intersection of AA and BB represents the solution corresponding to (14′) and (16′). An increase in μ will then shift the long-run equilibrium to Q' in the CR model (where F increases) and to Q'' in the alternative one (where F decreases). Unlike the CR model, where total assets $(m + F)$ are constant, we must have a reduction of assets in the alternative LRPF model.

References

Brock, William A. "Money and Growth: The Case of Long Run Perfect Foresight." *Internat. Econ. Rev.* 15 (October 1974): 750–77.

Calvo, Guillermo A., and Rodríguez, Carlos Alfredo. "A Model of Exchange Rate Determination under Currency Substitution and Rational Expectations." *J.P.E.* 85, no. 3 (June 1977): 617–25.

Sidrauski, Miguel. "Rational Choice and Patterns of Growth in a Monetary Economy." *A.E.R. Papers and Proc.* 57 (May 1967): 534–44.

Economies of Scale and Barriers to Entry

Richard Schmalensee

Massachusetts Institute of Technology

Dixit has recently presented a model in which established firms select capacity to discourage entry but cannot employ threats they would not rationally execute after entry. Entry deterrence in a slight modification of this model involves the classic limit-price output. Under linear or concave demand, however, the capital cost of a firm of minimum efficient scale is an upper bound on the present value of the monopoly profit stream that can be shielded from entry. It is argued that this suggests the general unimportance of entry barriers erected by scale economies.

In his seminal work on conditions of entry, Bain (1956) argued that the necessity for a firm to be large relative to the market in order to attain productive efficiency created a barrier to entry. The notion that large-scale entry can create a discrete difference between preentry and postentry price and profit levels is clear enough, but the details of Bain's argument are still somewhat controversial. In assessing the importance of the scale-economies barrier, Bain introduced the limit-pricing model of entry deterrence, in which established firms act as a perfect cartel and potential entrants expect those firms to maintain their preentry levels of output even after entry. This model has been subjected to strong criticism, however, in large part because it may not be rational for the established firms to keep output constant after large-scale entry has occurred.[1] Moreover, Stigler (1968) and others have challenged the basic idea that scale economies can create a

I am indebted to Steve Salop, Roger Bohn, and Nalin Kulatilaka for helpful comments, though I cannot share blame for remaining defects with them.

[1] Scherer (1980, chap. 8) provides a useful discussion of the limit-pricing model and its critics.

[*Journal of Political Economy*, 1981, vol. 89, no. 6]
© 1981 by The University of Chicago 0022-3808/81/8906-0010$01.50

meaningful entry barrier. These critics have stressed the fact that once an entrant has invested in an efficient plant, there is no difference (under the usual assumptions) between its position and that of established firms. Without a postentry difference, they have argued, there can be no real barrier to entry.

In part this last argument concerns the most useful definition of "barrier to entry." Spence (1977), building on insights of Schelling (1960), has recently presented an interesting analysis of the economic issues involved in that argument.[2] The basic point that emerges is that established firms, assuming they can coordinate their actions, have the advantage of being able to make some irreversible decisions before new entry appears. In particular, they can select the level of capacity facing any new entrant. Even if entry occurs and the established firms then wish to have less capacity, their preentry commitments may make a rapid reduction in capacity impossible. Recognizing this, entrants may be deterred. In his formal analysis, Spence assumes that the established firms build enough capacity to produce nearly the competitive output. Before entry, they produce the monopoly output, but they threaten to use all their capacity if entry occurs. If this threat is believed, as Spence assumes it is, entry can clearly be deterred. But a threat to increase output after entry is surely no more credible than the threat to maintain output that is the core of the limit-price model.

In an important recent paper, Dixit (1980) explores the implications of restricting the established firms to threats that they would find it in their postentry interests to execute. In particular, he assumes that potential entrants expect the postentry market equilibrium to be that of Cournot doupoly. That is, they expect the established firms to behave rationally but noncooperatively in response to their entry. Though this market model is far from perfect, it at least implies plausible profit-seeking behavior by both established sellers and new entrants, and it avoids building ad hoc behavioral asymmetries into the postentry situation.[3] Dixit finds that the established firms may still be able profitably to deter entry if they can commit to a level of capacity before potential entrants appear.

In Dixit's simplest model, scale economies arise because of fixed costs. In order to relate his analysis to the earlier limit-pricing literature and to empirical work on scale economies, it is assumed here instead that the long-run average cost curve is L-shaped, so that the minimum efficient scale of operations is well defined.[4] The resultant

[2] See also Dixit (1979), Spence (1979), and for a good discussion of the basic concepts involved, Salop (1979).

[3] In addition to the discussion of the Cournot assumption by Dixit (1980), see Orr and MacAvoy (1965), where it is argued that Cournot behavior is a plausible threatened reaction to cheating on a cartel agreement

[4] See, e.g., Modigliani 1958 and Scherer 1980, chaps 4 and 8.

model is analyzed in the next section. It is shown there that entry deterrence involves charging exactly the classic limit price before entry and holding output constant if entry occurs. A basic difference between this analysis and that of Bain-Sylos, however, is that the threat to hold output constant after entry is not always credible. It is shown that if the threat is credible, and if the market inverse demand function is linear or concave, the preentry present value of excess profits that can be shielded from entry cannot exceed the capital (start-up) cost of a firm of minimum efficient scale. In Section II, it is argued that this upper bound and widely accepted estimates of the importance of scale economies imply that the entry barriers considered here are generally of little quantitative importance. The implications of localized competition and long-lived advertising in this context are discussed briefly.

I. Credible Entry Deterrence

Following most of the relevant literature, the established firms are assumed to be either a single enterprise or a perfect cartel and are referred to as "the monopoly" in what follows. Only one potential entrant, called "the entrant," is considered. The monopoly is assumed to select and construct its capacity, K^m, prior to the entrant's appearance. Both the monopoly and the entrant operate under the cost conditions shown in figure 1. Economies of scale are such that firms with capacity less than K_0 have prohibitively high costs, so that K_0 is the minimum efficient scale in this industry. For levels of output, q, between K_0 and capacity, K, variable cost is v per unit. Capacity costs c per unit to install. It is assumed to last forever, so that the flow cost of capacity per period is (rc) per unit, where r is the relevant discount rate. With capacity K in place, the heavy kinked line in figure 1 is the

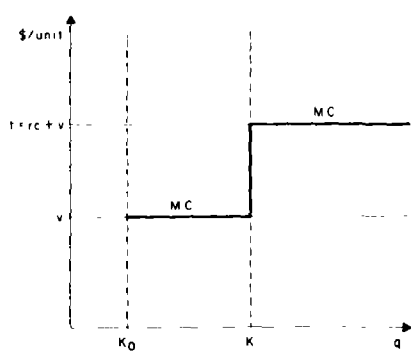


FIG. 1.—Marginal cost after acquisition of capacity K

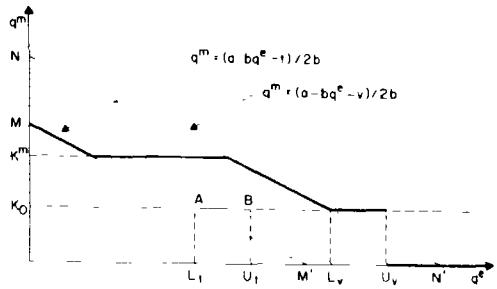


FIG. 2 — The monopoly's postentry reaction function

effective marginal cost schedule. By selecting K^m, the monopoly determines where the kink in its marginal cost function occurs. The entrant, with no sunk investment in capacity, faces a constant marginal cost of $t = rc + v$ for all output levels above K_0. All this follows the basic setup of Dixit's (1980) simplest model, except that he has $K_0 = 0$, and he allows nonzero fixed setup costs.

The first step in the analysis is the description of the monopoly's postentry Cournot reaction function. For illustrative purposes, suppose that the industry inverse demand function is given by

$$P = P(Q) = a - bQ = a - b(q^e + q^m), \tag{1}$$

where a and b are positive constants, P is price per unit, Q is total industry output, q^m is monopoly output, and q^e is the entrant's output. If marginal cost were some constant, X, the monopoly's reaction function would be simply

$$q^m = (a - bq^e - X)/2b. \tag{2}$$

Things are more complex under the cost structure assumed here. The heavy kinked line in figure 2 shows the monopoly's postentry reaction function given installed capacity K^m and market demand given by (1).

In figure 2, the line MM' is the reaction function of a firm choosing q^m, taking q^e as fixed, with constant marginal cost equal to t, while NN' is drawn for marginal cost v. If q^e exceeds U_v, the addition of K_0 to industry output would drive price below average variable cost, v, so that the monopoly's optimal output is zero even with capacity in place. For q^e between L_v and U_v, the monopoly optimally sets $q^m = K_0$. If there were no lower bound constraint on output, a smaller value of q^m would be chosen along NN'. As q^e is reduced below L_v, the monopoly expands output along NN' until the capacity constraint is encountered. The horizontal section of the reaction function in figure 2 corresponds to the vertical portion of the marginal cost schedule in

figure 1. If q^e is sufficiently low, it may (depending on the initial choice of K^m) be optimal for the monopoly to add capacity and expand output, at a marginal cost of t per unit. As q^e goes to zero, the monopoly's optimal output goes to M, the unconstrained monopoly level, as figure 2 is drawn. If K^m were chosen above M but below N, q^m would be set equal to K^m for $q^e = 0$. Monopoly output never exceeds N.

If the monopoly had not acquired any capacity, it would not invest if q^e were above U_t, since the addition of K_0 to industry output would drive the price below total unit cost, t. For q^e between L_t and U_t, it would be optimal to set $q^m = K_0$, while for smaller values of q^e it would be optimal to acquire more than the minimum possible capacity. Formally, the key quantity levels discussed above are given in general and in the linear case of equation (1) by

$$P(L_t + K_0) + K_0 P_Q(L_t + K_0) = t; \qquad L_t = (a - 2bK_0 - t)/b, \quad (3a)$$

$$P(U_t + K_0) = t; \qquad U_t = (a - bK_0 - t)/b, \quad (3b)$$

$$P(L_v + K_0) + K_0 P_Q(L_v + K_0) = v; \qquad L_v = (a - 2bK_0 - v)/b, \quad (3c)$$

$$P(U_v + K_0) = v; \qquad U_v = (a - bK_0 - v)/b. \quad (3d)$$

The reaction function relevant to the entrant's decisions is given by the mirror image of the schedule MAB in figure 2. With no commitment to capacity, the entrant has a constant (long-run) marginal cost of t for output levels above K_0. Consideration of figure 2 then makes it clear that the entrant will be deterred if it expects the monopoly's postentry output to exceed U_t. From (3b), U_t is exactly the preentry output predicted by the classic Bain-Sylos limit-price theory, with $P(U_t)$ the corresponding preentry price.

Figure 3 depicts a situation in which entry is deterred by monopoly choice of a capacity level, K^m, above the limit quantity, U_t. (Note the reversal of axes between fig. 2 and fig. 3.) Given its precommitment to

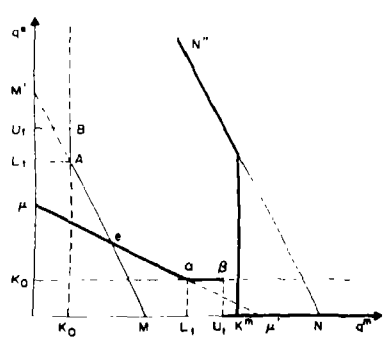


FIG. 3.—Credible entry deterrence

that capacity, the monopoly's reaction function is exactly as before; as long as q^e does not greatly exceed K_0, it is clear that the monopoly's optimal response is to utilize capacity fully. The entrant's reaction function, $\mu\alpha\beta$, is the mirror image of the ex ante monopoly schedule MAB. Note that with the monopoly's capacity in place and the corresponding costs sunk, the conventional Cournot equilibrium point, e, has no particular significance. The only equilibrium is $q^m = K^m, q^e = 0$; entry is deterred, and the monopoly always uses its installed capacity fully.

A bit more analysis of the situation depicted in figure 3 makes it clear that in order for the monopoly to be able to deter entry and still enjoy positive economic profit, two conditions are necessary. First, c must be positive; it must be possible for the monopoly to commit itself in advance to capacity, to purchase a lower marginal cost over some output range. If $c = 0$, the $MC = v$ reaction schedule $N''N$ (which corresponds to the left-hand portion of NN' in fig. 2) collapses to $M'M$, and the monopoly cannot avoid the Cournot point, e, which involves entry. Second, K_0 must be positive; there must be scale economies as well as durable capital. If $K_0 = 0$, the entrant's reaction function extends to the point μ', and K^m must exceed μ' in order to deter entry. But an examination of the entrant's reaction function makes it clear that $P(\mu') = t$, so that the monopoly can deter entry only by producing at least the competitive output in the absence of scale economies. (When $q^e = 0$, the entrant's marginal revenue is exactly the market price. Reference to [3b] makes it clear that $U_t < \mu'$, as fig. 3 shows.)

A somewhat stronger necessary condition for entry deterrence with positive profits is that the point β in figure 3 must lie inside the $N''N$ locus. That is, if marginal cost is v, it must be optimal for the monopoly to produce at least U_t in response to $q^e = K_0$. If not, any entrant can plan on producing K_0 and earning positive profit, regardless of the monopoly's choice of capacity or preentry output. If the market marginal revenue curve slopes downward, this condition means that the monopoly's marginal revenue at the point β must be at least v:

$$P(U_t + K_0) + U_t P_Q(U_t + K_0) \geq v. \tag{4}$$

Rearranging, noting that $P(U_t + K_0) = t$ from (3b), and multiplying by K_0, one obtains

$$[-P_Q(U_t + K_0)K_0]U_t \leq vK_0. \tag{5}$$

If the inverse demand function is either linear or concave, weak concavity implies

$$P(U_t) \leq P(U_t + K_0) + P_Q(U_t + K_0)[U_t - (U_t + K_0)]. \tag{6}$$

Substitution from (6) for the bracketed term in (5) yields

$$[P(U_t) - t]U_t \leq rcK_0. \tag{7}$$

Since U_t is the monopolist's minimum (preentry and postentry) deterrence output, condition (7) says that the flow of excess profit enjoyed by a monopoly that has credibly deterred entry cannot exceed the per-period cost of capital employed in an enterprise of minimum efficient scale.[5] This inequality clearly reflects the fact that both c and K_0 must be positive for this sort of limit pricing to be credible and profitable.

It should be noted that it may not be optimal for the monopoly to install sufficient capacity to deter entry even if it is possible to do so. Depending on the details of the demand structure, the monopoly might prefer to allow entry and have the equilibrium occur at some point along $e\alpha\beta$. It can, of course, guarantee itself any such point within $N'N$ by its choice of K^m. If the monopoly does decide to admit the entrant and the latter invests in capacity, inequality (7) serves to bound the total excess profit that the resultant duopoly can protect from entry by coordinated action.[6] In general, if entry is unattractive to outsiders, the flow of monopoly profits received by cooperating insiders cannot exceed the flow cost of capital assets embodied in a firm of minimum efficient scale, as long as industry demand is not strictly convex.

II. Extensions and Implications

The derivation of (7) rests on a number of strong assumptions that serve to overstate the ease of deterrence. First, it is at least arguable that the Cournot model overstates the intensity of rivalry to be expected in

[5] In this model, there would seem to be no general way to obtain a bound on profit, like condition (7), without using concavity. (If demand elasticity is a constant, $-E < -1$, [4] is satisfied for all positive K_0 if and only if $v \leq rc[E - 1]$.) Similar bounds hold in some related models of entry deterrence. In the "simple illustrative formal model" of Schmalensee (1978, pp. 312–13), e.g., if new brands do not expand total demand, per-brand excess profit in deterrence equilibrium equals brand-specific fixed cost. With output expansion, fixed cost becomes an upper bound. See also the equilibria in the references cited in Schmalensee (1978, p. 313, n. 15).

[6] If a given level of capacity is divided among several firms, each will see a larger Cournot marginal revenue at capacity output than would a single seller. For any given entrant output, capacity is thus more likely to be fully utilized with several established firms than with a single monopolist. This means that a higher level of profits can in principle be protected against entry by Cournot behavior than the monopoly analysis in the text implies. Given the difficulty of coordinating investment policies among several firms, however, there seems little to be gained by a detailed analysis. It is surely reasonable, though not necessarily correct, to assume that monopoly entry deterrence involves higher preentry profits than feasible oligopoly deterrence strategies.

the postentry duopoly.[7] If the entrant is more optimistic, entry will be harder to deter, and the flow of excess profits that can be protected will be smaller than (7) allows. Second, it was assumed that the monopoly was able to invest in infinitely durable assets. The recent work of Eaton and Lipsey (1980) supports the intuitive argument that the strength of the monopoly's commitment is reduced as the assets it purchases become less durable, since the monopoly cannot generally commit itself in advance to invest after entry in order to maintain capacity. Third, the assumption that average cost is effectively infinite for outputs below K_0 is very unrealistic. As the presentations of Modigliani (1958) and Scherer (1980, chap. 8) show, entry deterrence in the Bain-Sylos limit-pricing model is more difficult the more slowly average cost rises when output is reduced below K_0. The need to deter entry at small scale lowers the profitability of the limit-price point, which is exactly the point considered in deriving inequality (7). Finally, the assumption that the established firms manage to coordinate their investment policies perfectly in the interest of entry deterrence is very strong indeed. Less than perfect coordination in deterrence would be expected to make entry more likely.

Given the strength of these four assumptions, it would seem sensible to use (7) as an upper bound on monopoly profit associated with scale-economy entry barriers even in situations where industry inverse demand may be convex. To explore the implications of this conclusion, let us consider an industry with (price-level adjusted, net) assets A, earning an accounting rate of return r^* on those assets. Assuming away measurement problems and windfalls, accounting profits equal normal returns to capital plus monopoly profits. Let the competitive rate of return, the ratio of normal profits to A, be r. It is then sensible to use r to go from stocks to flows in this industry as is done in (7). That condition can then be rewritten as

$$(r^* - r)A \leq rcK_0, \tag{7'}$$

which immediately becomes

$$(r^* - r)/r \leq cK_0/A. \tag{8}$$

Assuming that economies of scale are essentially exhausted at capacity K_0, the ratio cK_0/A should be a good approximation to the ratio K_0/Q, the ratio of minimum efficient scale (in output terms) to market demand. For most industries, most scholars estimate this ratio to be less than 0.10. (Scherer [1980, chap. 4] presents a large number of such estimates.) Condition (8) then implies that if an industry has a return on assets of, say, 10 percent, entry is not taking place, and only

[7] See the discussion of that model in Scherer (1980, chap. 5)

scale economies can serve to deter entry, the normal rate of return must be at least 9.1 percent unless scale economies are exceptionally important. Scale-economy entry barriers thus cannot account for a rate of monopoly profit of even 1 percent, surely not much above the standard deviation of the corresponding measurement error in most situations.

An analysis of deadweight loss similarly suggests that entry barriers derived from scale economies are not generally important. If industry demand is given by (1), it is easy to show that the deadweight loss from limit pricing is $b(K_0)^2/2$. Letting demand elasticity, P/bQ, equal E, the ratio of deadweight loss to sales revenue becomes $[(K_0/Q)^2/2E]$. This exceeds 0.005 only if economies of scale are of exceptional importance or if E is less than one.

The analysis so far, based on conventional wisdom about the importance of scale economies, supports Bain's (1956, p. 212) conclusion that "economies of scale in production and distribution do not loom large as the basis of barriers to entry." Two factors that might weaken this conclusion deserve brief mention, however.

First, most U.S. estimates of the importance of scale economies assume national markets and implicitly treat products as homogeneous. If transport costs are important and markets are regional, scale economies can become very important indeed. Scherer (1980, p. 98) estimates that a cement plant of minimum efficient scale would account for about 40 percent of demand in a typical regional market, for instance. If products are differentiated, competition may be similarly localized in product space. As is discussed at length in Schmalensee (1978), this sort of demand structure also serves to magnify the importance of scale economies in entry deterrence.

Second, spending on advertising generally has an investment component, though there is considerable uncertainty about the typical durability of the corresponding asset.[8] The analysis of Section I above implies that, by itself, the longevity of advertising's effects on demand does not make it possible to use advertising to protect excess profits. A number of authors have suggested that the use of advertising involves significant economies of scale, however. There is also considerable uncertainty about the importance of these scale effects.[9] If it turns out in general or in some industry that scale economies in advertising are important *and* that advertising has relatively long-lasting effects on demand, the analysis above would lead one to suspect, by analogy,

[8] Comanor and Wilson (1979) survey the evidence on this point.

[9] See Schmalensee 1972, chap. 7; Comanor and Wilson 1979. Spence (1980) provides a useful discussion of the definition of economies of scale in this context, though he does not treat advertising as an investment.

that advertising could be used like investment in plant and equipment to protect excess profits from entry. In order to verify this suspicion, however, one would at least have to verify that at any instant past investment in advertising has the same sort of impact on *marginal* returns to changes in output that investment in capacity does. The first of the models discussed in Schmalensee (1974) illustrates that advertising does not have such effects in at least some plausible dynamic models.[10] Rather than attempting to reason by analogy from the treatment of investment in capacity here, it would seem preferable to construct explicit models involving economies of scale in advertising and durability of advertising's effects on demand and to see if those models allow established monopolies to deter entry by means of investment in advertising. Such models may even imply simple analogs of (7), but that remains to be seen.

References

Bain, Joe S. *Barriers to New Competition, Their Character and Consequences in Manufacturing Industries.* Cambridge, Mass. Harvard Univ. Press, 1956.

Comanor, William S., and Wilson, Thomas A. "The Effect of Advertising on Competition: A Survey." *J. Econ. Literature* 17 (June 1979): 453–76.

Dixit, Avinash. "A Model of Duopoly Suggesting a Theory of Entry Barriers." *Bell J. Econ.* 10 (Spring 1979): 20–32.

———. "The Role of Investment in Entry-Deterrence." *Econ. J.* 90 (March 1980): 95–106.

Eaton, B. Curtis, and Lipsey, Richard G. "Exit Barriers Are Entry Barriers: The Durability of Capital as a Barrier to Entry." *Bell J. Econ.* 11 (Autumn 1980): 721–29.

Little, John D. C. "Aggregate Advertising Models: The State of the Art." *Operations Res.* 27 (July–August 1979): 629–67.

Modigliani, Franco. "New Developments on the Oligopoly Front." *J.P.E.* 66, no. 3 (June 1958): 215–32.

Orr, Daniel, and MacAvoy, Paul W. "Price Strategies to Promote Cartel Stability." *Economica* 32 (May 1965): 186–97.

Salop, Steven C. "Strategic Entry Deterrence." *A.E.R. Papers and Proc.* 69 (May 1979): 335–38.

Schelling, Thomas C. *The Strategy of Conflict.* Cambridge, Mass.: Harvard Univ. Press, 1960.

Scherer, Frederic M. *Industrial Market Structure and Economic Performance.* 2d ed. Chicago: Rand McNally, 1980.

[10] But see also the "alternative model" discussed briefly in Schmalensee (1974, pp. 586–87) and the treatment of advertising in Spence (1977). To the extent that a certain minimum amount of introductory advertising is necessary to make any sales at all, such introductory outlays would seem to have the required impact on marginal returns. In Schmalensee (1978), such introductory advertising is treated as a lumpy investment central to entry deterrence. Little (1979) provides an illuminating survey of dynamic advertising/sales models.

Schmalensee, Richard. *The Economics of Advertising*. Amsterdam: North-Holland, 1972.

———. "Brand Loyalty and Barriers to Entry." *Southern Econ. J.* 40 (April 1974): 579–88.

———. "Entry Deterrence in the Ready-to-Eat Breakfast Cereal Industry." *Bell J. Econ.* 9 (Autumn 1978). 305–27.

Spence, A. Michael. "Entry, Capacity, Investment and Oligopolistic Pricing." *Bell J. Econ* 8 (Autumn 1977): 534–44.

———. "Investment Strategy and Growth in a New Market." *Bell J. Econ.* 10 (Spring 1979): 1–19.

——— "Notes on Advertising, Economies of Scale, and Entry Barriers." *Q.J.E.* 95 (November 1980): 493–508.

Stigler, George J. "Barriers to Entry, Economies of Scale, and Firm Size." In *The Organization of Industry*. Homewood, Ill.: Irwin, 1968.

The Great Depression and Commodity-exporting LDCs: The Case of Brazil

Eliana A. Cardoso

Boston University

This paper discusses the role of fiscal and monetary policies in an LDC with a major commodity export that is facing a depression abroad. The analysis is conducted in the context of a general equilibrium model which comprises a commodity-producing sector and an import-competing sector. Idle capacity and labor unemployment are assumed. Flexibility of the real exchange rate and industrial output responses to demand are the main adjusting mechanisms. The key to understanding the behavior of the Brazilian economy in the 1930s is the government coffee-support policy, which held the income of the export sector at a high level and hence enabled the manufacturing sector to expand. The fact that industrial output expanded in the face of both real depreciation and appreciation points to the importance of expenditure effects relative to price effects during the 1930s.

This paper discusses the role of fiscal and monetary policies in an LDC with a major commodity export facing a depression abroad. The analysis is conducted in the context of a general equilibrium model which comprises a commodity-producing sector and an import-competing sector that produces both consumption and investment goods. Idle capacity and labor unemployment are assumed. Flexibility of the real exchange rate and industrial output response to demand are the main adjusting mechanisms.

I am indebted to R. Dornbusch and to several anonymous referees for their comments and suggestions. Financial support was provided by a grant from the Ford Foundation.

[*Journal of Political Economy*, 1981, vol. 89, no. 6]
© 1981 by The University of Chicago. 0022-3808/81/8906-0004$01.50

Having presented the model in the first section, we use it as the basis of our discussion of industrialization in Brazil in the years of the Great Depression. The key to understanding the behavior of the Brazilian economy in the 1930s is the government coffee-support policy, which held the income of the export sector at a high level and hence enabled the manufacturing sector to expand.[1] We show that stockpiling expanded income, and that, when stockpiling was financed by export duties, the trade balance improved, but when it was financed by credit, a trade deficit arose. The fact that the currency was devalued during periods in which the coffee sector was supported through domestic credit expansion but appreciated when coffee purchases were financed via duties strengthens the hypothesis that devaluations were a response to the external disequilibrium created by the coffee policy. The fact that industrial output expanded in the face of both real depreciation and appreciation points to the importance of expenditures effects, relative to depreciation and price effects, induced by fiscal and monetary policies during the 1930s.

I. The Model

The economy has two sectors: a coffee-producing sector oriented to the export market and a manufacturing sector that produces goods for domestic consumption and investment. Imported goods compete with domestic manufactures on the internal market.

There are markets for three types of goods: coffee (C), domestic manufactures (Q), and imported manufactures (M). The prices of these goods in home currency are, respectively, P_c, P, and P_m. The nominal wage is denoted W. The world supply of imported manufactures is perfectly elastic at the international price P_m^*, so that $P_m = eP_m^*$, where e stands for the exchange rate.

When imported manufactures are used as the numeraire, the *real*

[1] This hypothesis is advanced by Furtado (1963). Furtado is criticized by Pelaéz (1972), who contends that a large share of the income generated by the multiplier effects of the coffee-support expenditures was canceled by the offsetting negative multiplier effects of the taxes levied to finance coffee purchases. Fishlow (1972) argues that the coffee-support program generated demand in a somewhat more complex manner than Furtado believes. Basing his argument on a partial equilibrium model, Fishlow maintains that the greater part of the export duties were passed on to the foreign consumer. The transfer of the duties was guaranteed by the inelasticity of demand and the high elasticity of supply. Our more complete model incorporates the mechanism described by Fishlow, while at the same time analyzing the coffee policy from the standpoint of its impact on real income in accord with the approach advised by Pelaéz and Suzigan (1976). A monetary explanation for the industrial upsurges during the 1930s is found in Pelaéz (1972) and Neuhaus (1973). The monetary factors explaining the behavior of economic policy in the Brazilian experience during the Depression are also incorporated in our model.

prices are: $p_c \equiv P_c/P_m$, $p \equiv P/P_m$, and $w \equiv W/P_m$. We assume unemployment and a given nominal wage rate.

A. *The Coffee Market*

In the coffee sector, the fixed factor (suitable land planted in coffee and served by adequate transport) can be combined with varying quantities of labor to obtain the crop. The grower maximizes profits by equalizing the value of the marginal product and a given nominal wage. When the price of coffee rises, so do employment and output. The coffee supply is, therefore, a rising function of the price of coffee in wage units:

$$C^s = C^s(p_c/w). \tag{1}$$

Coffee is exported. Foreign demand is a function of the real income of the coffee-importing countries, y^*, and of the dollar price of coffee deflated by the world price of imported manufactures:

$$C^D = C^D(Tp_c, y^*), \tag{2}$$

where $p_c = (P_c/eP_m^*)$ and $T \equiv 1 + t$, with t being the duty rate on coffee exports and P_c and P_m^* the home price of coffee and foreign currency price of manufactures.

Part of the coffee is bought by the government rather than sold abroad. These purchases are financed through foreign loans, domestic credit, and export duties on coffee. The real coffee purchases on the part of the government are designated G.

Equilibrium on the coffee market requires that[2]

$$C^s(p_c/w) = C^D(p_cT, y^*) + G. \tag{3}$$

Equation (3) determines the real producers' price of coffee, p_c, as a function of the exogenous variables:[3] $p_c = p_c(y^*, G, T, w)$: $\partial p_c/\partial y^* > 0$; $\partial p_c/\partial G > 0$; $\partial p_c/\partial T < 0$; and $\partial p_c/\partial w > 0$.

We also define real export earnings, X, and real disposable income of the coffee sector, y_c, measured in terms of imported manufactures, respectively, as:[4]

[2] It is assumed that domestic coffee consumption is negligible.

[3] From (3).

$$p_c' = \frac{1}{\epsilon_c + \eta_c}(\epsilon_c w' + \eta_{y^*} y^{*\prime} + \sigma - \eta_c T'), \tag{4}$$

where a prime beside the variable indicates its logarithmic derivative, i.e., $x' = dx/x$; $\sigma = dG/C^d$; ϵ_c = price elasticity of the coffee supply, η_c = absolute value of the price elasticity of the demand for coffee, and η_{y^*} = income elasticity of the demand for coffee. Equation (4) is an approximation. It is exactly correct only if initially $G = 0$ or $C^s = C^d$.

[4] The effect of changes in the exogenous variables on real export earnings can be

$$X \equiv Tp_cC^D(Tp_c, y^*) = X(y^*, G, T, w), \tag{7}$$

$$y_c \equiv p_cC^s(p_c/w) = y_c(y^*, G, T, w). \tag{8}$$

Real export earnings and real disposable income of the coffee sector are increasing functions of foreign income, government purchases, and real wages. A rise in export taxes raises export earnings but reduces disposable income.

An increase in the real income of the coffee-importing countries, y^*, enhances external coffee demand and raises the real price of coffee, real export earnings, and real disposable income of the coffee sector.

An increase in government coffee purchases boosts the price, and, since the demand for coffee is inelastic, export earnings and real income of the coffee sector rise.

An increase in export duties, T, raises the dollar price and reduces international demand for coffee, bringing on a decrease of the real price received by the grower. Since the demand for coffee is inelastic relative to its price, export earnings rise. But disposable income of the coffee sector falls, because both the quantity produced and the price received by the grower diminish.

An increase in the real wage rate lessens coffee supply and pushes the real price of coffee up. Both export earnings and real income rise inasmuch as demand is inelastic.

B. *The Market for Domestic Manufactures*

The price of domestic manufactures is cost determined. The real price of domestic manufactures, in terms of imported manufactures, is

$$p = \beta w, \tag{9}$$

where β stands for the output/labor ratio.

Manufacturing output is determined by demand, which depends on real domestic income, y, on its price in relation to that of imported similars, p, and on real balances, h. Real domestic manufacturing output, y_q, thus is:

$$y_q = pQ(p, y, h). \tag{10}$$

---

studied using the following equation:

$$X' = \left(\frac{1 - \eta_c}{\epsilon_c + \eta_c}\right)(\epsilon_c w' + \sigma + \epsilon_c T') + \left(\frac{1 + \epsilon_c}{\epsilon_c + \eta_c}\right)\eta_{y^*}y^{*\prime}. \tag{5}$$

The effect of changes in the exogenous variables on the income of the coffee sector is given by:

$$y_c' = \left(\frac{1 + \epsilon_c}{\epsilon_c + \eta_c}\right)(\eta_{y^*}y^{*\prime} + \sigma - \eta_c T') + \left(\frac{\epsilon_c}{\epsilon_c + \eta_c}\right)(1 - \eta_c)w'. \tag{6}$$

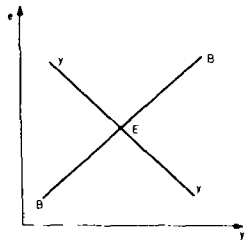


FIG. 1

Real income is defined as:

$$y \equiv y_c + y_q. \tag{11}$$

Equations (9), (10), and (11) allow us to determine internal equilibrium and to express real income as a function of the income of the coffee sector, real balances, and the real wage.[5] Given the money stock, the fiscal policy, and the money wage, we can depict internal equilibrium by an inverse relationship between real income and the exchange rate, as illustrated in figure 1 by schedule yy.[6] Assume an initial combination of the real income and the exchange rate for which there is internal equilibrium. An increase in the exchange rate reduces the real wage, real balances, and income of the coffee sector. Both the reduction in real balances and the reduction in the income of the coffee sector contribute to reducing the demand for domestic goods. However, the reduction in real wages, and thus in the price of domestic goods in relation to imports, increases demand for domestic manufactures in substitution for the imported goods. As long as the two first effects dominate, a depreciation of the exchange rate reduces demand for domestic manufactures. Consequently, industrial output

[5] Assuming that, initially, $G = t = 0$ and $X = M$, where M stands for imports, we have $y_c = X - M$ and, from eq. (11),

$$dy/M = y_c' + (y_q/M)y_q'. \tag{12}$$

From eq. (10),

$$y_q' = (1 - \eta_q)p' + q(dy/y_q) + \lambda_q h', \tag{13}$$

where: η_q = absolute value of the price elasticity of the demand for domestic manufactures, q = marginal propensity to consume domestic manufactures, and λ_q = elasticity of demand for domestic manufactures in relation to real balance. Substituting (13) into (12) and observing, from the budget constraint, that $(y_q/M)(1 - \eta_q) = \eta_m$, where η_m is the price elasticity of the demand for imported manufactures, we obtain

$$dy/M = [1/(1 - q)](y_c' - \eta_m p' + \Theta\lambda_q h'), \tag{14}$$

where $\Theta \equiv y_q/M$.

[6] Note that $h' = p' + w = -e'$, for given nominal money stock and nominal rate. For given policy, we also have: $y_c' = \phi_c w'$, where $\phi_c = (1 - \eta_c)[\epsilon_c/(\epsilon_c + \eta_c)]$. We can thus rewrite (14) as $e'/(dy/M) = -(s + m)/(\phi_c + \Theta\lambda_q - \eta_m)$, where s = the marginal propensity to save. The schedule yy is downward sloping as long as $(\phi_c + \Theta\lambda_q) > \eta_m$.

falls, and real income is reduced. It is implied that a depreciation is contractionary when the nominal money supply is held constant.

C. *The Balance of Payments*

Imports of manufactures, M, are a function of their price relative to that of domestic similar products, of real income, and of real cash balances:

$$M = M(p, y, h). \tag{15}$$

The balance of payments measured in terms of imported manufactures is

$$B = X - M + F, \tag{16}$$

where F stands for independent movements of foreign capital expressed in terms of imported goods.

External equilibrium, $B = 0$, can be represented by a positive relationship between the exchange rate and real income, as illustrated by schedule BB in figure 1.[7] If real income increases, the demand for imports rises, generating a trade deficit. To restore equilibrium, the exchange rate has to depreciate. For the depreciation to improve the balance of payments, the reduction in coffee-export revenues has to be smaller than the reduction in import spending, obtained through the substitution effect, and the reduction in cash balances, induced by the depreciation.

As illustrated in figure 1, internal and external equilibrium prevails at E, where yy and BB cross. We shall now illustrate the model by presenting some historically relevant comparative static exercises.

D. *Coffee Purchases Financed by Export Duties*

Assuming that $G = t = 0$, at the outset, the result of the government's increasing expenditures while maintaining a balanced budget ($dG = dtC^D$) is as follows.

Since the demand for coffee is inelastic, the initial impact of government outlays financed by export duties is to raise coffee prices and export earnings,[8] creating a trade surplus and shifting BB to the right,[9] as in figure 2. On the other hand, as the income of the coffee

[7] To obtain the slope of BB, we assume $B = 0$, $X = M$, and $F = 0$, at the outset. We thus differentiate (16) totally to obtain $e'/(dy/M) = m/(\eta_m + \lambda_m - \phi_c)$: m = the marginal propensity to spend in imported manufactures; and λ_m = the elasticity of demand for imports in relation to real balances.

[8] Observe that $dX = dy_c = \{(1 + \epsilon_c)[(1 - \eta_c)/(\epsilon_c + \eta_c)]\}P_c dG$.

[9] The schedule BB shifts to the right in the following proportion: $dy = (1/m)dX$.

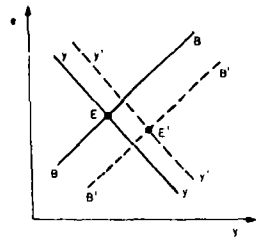


FIG. 2

sector rises, the demand for manufactures is promoted, and industrial production expands (yy shifts to the right).[10] At the higher income level, there still is a trade surplus.[11] The higher the marginal propensity to save, the larger is the balance-of-payments surplus, because less expenditures on foreign manufactures will be induced by the higher domestic income. The exchange rate appreciates to correct for the external surplus.

We thus conclude that coffee purchases financed by export duties are expansionary and induced an appreciation of the exchange rate.

E. Coffee Purchases Financed by Foreign Loans

Next, we consider coffee purchases financed by foreign loans. As in the preceding case, increases occur in the price of coffee, in export earnings, and in income of the coffee sector. Likewise, industrial output expands, and domestic income grows. With the increase in imports, a trade deficit lower than the government deficit arises. Since the government deficit is offset by foreign loans, the balance of payments improves. Coffee purchases financed by foreign loans are, thus, expansionary and induce an appreciation of the exchange rate, ceteris paribus.[12]

F. Coffee Purchases Financed by Printing Money

Consider, now, coffee purchases financed by printing money ($p_c dG = dh$). As already seen, the increase in government expenditures on

[10] The schedule yy shifts to the right in the following proportion, $dy = [1/(s + m)]dy_c$.

[11] At the going exchange rate, the trade surplus can be measured by $dB = [s/(s + m)]dy_c$.

[12] Observe that, when coffee purchases are financed by foreign loans, $dX = [(1 - \eta_c)/(\epsilon_c + \eta_c)]dG$ and $dy_c = [(1 + \epsilon_c)/(\epsilon_c + \eta_c)]dG$. In terms of our figures, yy shifts to the right in the following proportion, $dy = [1/(s + m)]dy_c$, and BB shifts to the right in the following proportion, $dy = \{(1/m)[(1 + \epsilon_c)/(\epsilon_c + \eta_c)]\}dF = (1/m)dy_c$.

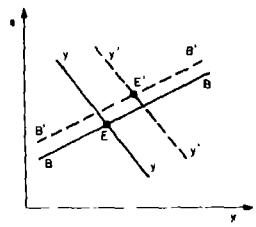


FIG. 3

coffee will raise the price of the product, export earnings, and the income of the coffee sector. Observe that the increase in the income of the coffee sector, for an increase in government expenditures on coffee, is larger than the increase in export revenues since it also grows by the government deficit.

The increase in money balances induces more spending on imports. If this effect exceeds the increase in export revenues, even at the initial income level, a trade deficit appears, shifting *BB* to the left,[13] as shown in figure 3. On the other hand, at the going exchange rate, the increases in real money balances and in the income of the coffee sector contribute to creating excess demand for domestic goods. Accordingly, the equilibrium output level of the manufacturing sector is enhanced, and real income grows.[14]

At the going exchange rate, a balance-of-payments deficit appears. The external disequilibrium is corrected as the exchange rate depreciates: Domestic goods substitute for imports, and the reduction in real balances reduces demand for both domestic and imported goods.

We thus conclude that financing coffee purchases through credit or duties has far different effects for the balance of payments. The government can also finance its coffee purchases through a combination of duties, credit, and foreign loans. The final effect will certainly be an expansion of income, but the effect on the exchange rate then depends on the shares of credit, export duties, and foreign loans in the total government expenditures on coffee, and on the savings propensity.

II. The 1930s: An Application

Figure 4 exhibits the Brazilian industrial output expansion during the 1930s in the face of alternating real depreciation and appreciation. The model developed in the previous section asserts that the coffee-

[13] The shift in *BB* can be measured by $dy = \{(1/m)[(1 - \eta_c)(\epsilon_c + \eta_m) - \lambda_m(M/h)]\}dh$

[14] The shift in *yy* is measured by $dy = \{[1/(s + m)][(1 + \epsilon_c)(\epsilon_c + \eta_c) + (y_g/h)\lambda_g]\}dh$

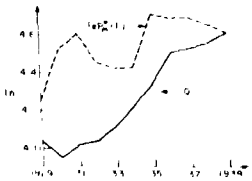


FIG. 4.—Real industrial output and the real exchange rate, Brazil, 1929–39. Source table 2.

TABLE 1

RATES OF DEPRECIATION IN BRAZIL, 1928–39

Period	$(e - e_{-1})/e_{-1}$
1928–31	.68
1931–33	−.11
1933–35	.37
1935–37	−.07
1937–39	.19

SOURCE.—Table 3.

support policy pursued during the 1930s in Brazil accounts for the industrial output expansion observed in the period. It also follows from the model that the observed appreciation rates, shown in table 1, can be attributed to the fact that, in the periods 1931–33 and 1935–37, stockpiling was mainly financed by export duties, and that the observed depreciations, in the years 1933–34 and 1937–39, can be attributed to the fact that, during those periods, coffee purchases were mainly financed by domestic credit. We further explore these facts in this section.

In the course of the 1920s, the coffee-support program proved quite effective. By purchasing unsalable stocks, the government held the international coffee price at high levels. Because of this artificial stimulus, coffee production nearly doubled from 1925 to 1929 at the same time that export remained approximately constant. In 1927–29, it was possible to export only two-thirds of the harvest.[15]

When the crisis erupted in 1929, both the federal government and the state of São Paulo abandoned their coffee-support programs because the possibility of external financing had vanished, and the Instituto Paulista do Café was in financial straits.[16] Between 1928 and 1930, the coffee price index in cruzeiros fell from 100 to 63.7.[17]

[15] See Furtado 1963, p. 198.
[16] See Silber 1977, p. 189.
[17] See Malan and Bonelli 1977, table 1.2.

These facts can be fitted into the model presented in Section I. With the relinquishment of the support policy and the drop in the coffee price, both export earnings and domestic income decrease. This, in turn, lowers industrial output (the industrial production index fell from 58 in 1928 to 52 in 1930 [base = 1939])[18] as well as causing expenditures on manufactured imports to decrease more than export earnings. Even so, the trade surplus[19] is outweighed, and the government is forced to devalue the exchange rate in order to deal with the balance-of-payments deficit.

In mid-1930, the São Paulo State government managed to arrange a foreign loan for the purpose of guaranteeing coffee prices,[20] though at this point the mechanism must be qualified because the year 1930 was marked by serious political disturbances and a withdrawal of capital that contributed to the deterioration of the balance of payments as well as to the devaluation of the exchange rate. By assuring minimum purchase prices that were profitable for the majority of the coffee growers, the program actually maintained the employment level of the export economy and, indirectly, of the production sectors linked to the domestic market. Thanks to the coffee policy, industrial production began to recover from its 1930 bottom.

In May 1931, the Conselho Nacional do Café was established to support the coffee sector by purchasing and destroying stocks. The purchases were financed through duties levied on coffee exports and funds obtained from the Banco do Brasil and the National Treasury. In 1931 and 1932, the coffee bought by the government accounted for roughly 30 percent of the export revenues. Between May 1931 and February 1933, 65 percent of these expenditures were financed by duties and the remaining 35 percent by credit from the Banco do Brasil and the National Treasury.[21]

As the income of the coffee sector recovered from its lowest levels thanks to the price-support policy, the increasing demand for manufactures led to a rise in industrial output. Since government expenditures were in part financed by duties, higher income levels were compatible with the improvement in the balance of trade and the appreciation of the exchange rate.

Between February 1933 and December 1934, the shares of the financial sources changed considerably, with domestic credit coming to assume a more important role. To better understand this policy, one might glance at figure 3, where the manufacturing sector expands in the presence of exchange depreciation.

[18] See Fishlow 1972, table A-1.

[19] Between 1928 and 1930, export earnings fell from US$473.4 million to US$319.4 million, and imports fell from US$383.3 million to US$225.1 million.

[20] See Silber 1977, p. 190.

[21] Ibid., p. 192.

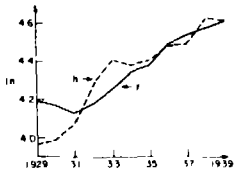


FIG. 5.—Real income and the real money stock, Brazil, 1929–39. Source: table 2.

TABLE 2

REAL OUTPUT, REAL MONEY, AND REAL EXCHANGE RATE INDICES FOR BRAZIL: 1928–39

Years	Index of Real Industrial Output Q (1)	Index of Real Output y (2)	Index of Money Balances Deflated by the GDP Deflator (3)	Index of Import Prices Deflated by Industrial Prices eP_m^*/P (4)
1928	58	65.2	51 5	67 2
1929	56	65 9	52 9	69 9
1930	52	64.5	54 1	91.7
1931	55	62.4	64 8	99.7
1932	56	65 1	72 1	86 2
1933	61	70 9	81 2	83.0
1934	68	77.4	80 2	84.0
1935	77	79 7	81 8	110.6
1936	91	89 3	88 0	110 3
1937	93	93 4	89 2	110 0
1938	96	97 6	101 9	105 5
1939	100	100.0	100 0	100 0

SOURCES.—Col. 1, Fishlow (1972), table A-1. Col. 2, Haddad (1977), pp. 11, 191. Cols. 3 and 4, table 3.

TABLE 3

MONEY STOCK AND PRICE INDICES FOR BRAZIL: 1928–29

Years	Money Stock Annual Average H (1)	Dollar Exchange Rate e (2)	Import Prices eP_m^* (3)	Industrial Prices P (4)	GDP Deflator (5)
1928	52.9	44 3	56.7	84.4	102.7
1929	52.2	44 3	54 6	78 1	98.7
1930	46.8	48.4	59.4	64 8	86.5
1931	49.9	74 5	69 2	69 4	77.0
1932	56.5	73 4	60 1	69 7	78 4
1933	62 2	66.1	59.4	71.6	76 6
1934	65.4	73.4	63.0	75 0	81 5
1935	70.0	90.6	85 7	77 5	85 6
1936	76 5	90 1	91 3	82 8	86 9
1937	84.8	83.9	94 8	86 2	95 1
1938	100 1	91 7	99.0	93 8	98 2
1939	100.0	100 0	100 0	100 0	100 0

SOURCES.—Col. 1, Pelaez and Suzigan (1976), table A-3. Col. 2, Malan and Bonelli (1977), table 1 3. Col. 3, ibid., table A V 12; this is a tariff-inclusive cruzeiro price of imports. Cols. 4 and 5, ibid., table A V 14.

Between 1935 and 1937, the Departamento Nacional do Café controlled the export supply through a quota system whereby coffee growers were obliged to turn part of their production over to the government for the formation of stocks. This was similar to financing coffee purchases through duties. During this phase there were less financial resources, and the manufacturing sector expanded in the presence of a slight exchange appreciation.

Figures 4 and 5 and tables 1, 2, and 3 summarize the main facts.

References

Fishlow, Albert. "Origins and Consequences of Import Substitution in Brazil." In *International Economics and Development: Essays in Honor of Raúl Prebisch*, edited by Luis Eugenio Di Marco. New York: Academic Press, 1972.

Furtado, Celso. *The Economic Growth of Brazil: A Survey from Colonial to Modern Times*. Translated by Ricardo W. Aguiar and Eric C. Drysdale. Berkeley and Los Angeles: Univ. California Press, 1963.

Haddad, Claudio. "Crescimento do Producto Real Brasileiro, 1900–1947." In *Formação Econômica do Brasil*, edited by Flavio R. Versiani. São Paulo: Edição Saravia, 1977.

Malan, Pedro, and Bonelli, Regis. *Política Econômica Externa e Industrialização Brasil (1939–1952)*. Rio de Janeiro: Inst. Pesquisa Econ. e Soc., 1977.

Neuhaus, Paulo. "The Monetary History of Brazil: 1900–1945." Mimeographed. Univ. Chicago, Dept. Econ., 1973.

Pelaéz, Carlos M. *História de Industrialização Brasileira: Crítica a Teoría Estruturalista do Brazil*. Rio de Janeiro: Análise e Perspectiva Econ., 1972.

Pelaéz, Carlos M., and Suzigan, Wilson. *História Monetária do Brasil: Análise da Política, Comportamento e Instituições Monetárias*. Rio de Janeiro: Inst. Pesquisa Econ. e Soc., 1976.

Silber, S. "Analise da Política Econômica e do Comportamento da Economia Brasileira durante o Período 1929–1939." In *Formação Econômica do Brasil*, edited by Flavio R. Versiani. São Paulo: Edição Saravia, 1977.

Confirmations and Contradictions

Estimating Property Tax Capitalization: A Further Comment

Raymond M. Reinhard

In recent years, an increasingly popular area of empirical investigation has concerned attempts to test Tiebout's (1956) hypothesis that allocative efficiency in the provision of public services may be attained through a system of local governments.[1] In his seminal contribution to the analysis of the Tiebout hypothesis and the capitalization of property taxes, Oates (1969) reasoned that, if consumers are sensitive to the public service benefits and property tax costs associated with home ownership, these factors should be reflected in housing prices. Until recently, most studies of the capitalization phenomenon have followed Oates's specification in which housing values are regressed on a vector of housing characteristics, public service measures, and a tax price term.[2]

As several analysts have noted,[3] however, attempts to "test" the Tiebout hypothesis by using an Oates-type equation must certainly lead to failure. This is so because, in a perfect Tiebout equilibrium, a marginal increase in housing value brought about by increased provision of public services must necessarily be offset by a marginal decrease in housing value caused by the added tax cost (assuming that

I am grateful to Lee S. Friedman, Paul R. Portney, Phillip E. Vincent, A. Thomas King, and an anonymous referee for their helpful comments. Wallace Oates, Jon C. Sonstelie, and Paul R. Portney kindly made their data available to me.

[1] See, e.g., Oates 1969, 1973; Pollakowski 1973; Edel and Sclar 1974; Hamilton 1976*a*, 1976*b*; Meadows 1976; King 1977; Rosen and Fullerton 1977; Epple, Zelenitz, and Visscher 1978; Linneman 1978; and Sonstelie and Portney 1980.

[2] A recent exception is Epple, Zelenitz, and Visscher (1978).

[3] See Edel and Sclar 1974; Hamilton 1976*b*; and Linneman 1978.

[*Journal of Political Economy*, 1981, vol. 89, no. 6]
© 1981 by The University of Chicago. 0022-3808/81/8906-0015$01.50

the public services are fully financed by property taxes). Thus, as Linneman (1978) recently observed, "[e]conometrically, an attempt to estimate [an Oates-type equation] will be rendered impossible due to data matrix singularity in taxes and some linear combination of the public goods" (p. 536).

The Contribution of King

In a recent issue of this Journal, A. Thomas King (1977) offered two significant contributions to the measurement of capitalization effects. First, he observed that most prior studies relied on an improper specification of the tax price term. As King noted, "[w]hereas the hypothesis suggests capitalization of the tax *burden*, this [Oates's] equation suggests capitalization based upon the tax *rate*" (p. 426). Second, and more importantly, King suggested that the tax price term be included as part of the dependent variable. His stated concern was with avoiding a possible bias in the coefficient of the tax price term; a more cogent rationale for the alternative specification is that it avoids the data matrix singularity problem mentioned earlier.[4]

Unfortunately, King's suggested estimation technique is flawed by an error of econometrics. A reanalysis of his results, summarized below, shows that King's technique, as originally presented, will lead to biased estimates of the true extent of capitalization. The econometric error, however, is easily remedied; as corrected, the King technique appears to offer a significant improvement over the standard Oates-type equation.

The Model

King suggests that an equation of the following form be used to estimate the extent of property tax capitalization:

$$V - btV = \beta_0 + \Sigma\beta_i X_i + \Sigma\beta_j P_j. \tag{1}$$

In equation (1), V represents the market value of the property, t is the property tax rate, and X_i and P_j represent physical attributes of the property and measured public services, respectively. The variable b measures the extent to which the property tax bill, tV, is capitalized into the property value, ". . . and is constrained to a value selected from the interval (0.1–1.0)" (King, p. 429). King suggests that "a maximum likelihood estimate of the extent of capitalization can be

[4] In addition, by incorporating the tax price term on the left-hand side of the estimating equation, the dependent variable then reflects the actual user cost of the residence, while the right-hand-side (independent) variables measure the services which flow from the house in a standard hedonic price equation framework.

obtained by varying b over the interval and reestimating, observing changes in R^2" (ibid.). Applying this methodology to Oates's data, King concludes that the result of the misspecification in the original Oates equations is to overestimate the true extent of property tax capitalization.

Before discussing the econometric error, two points deserve brief mention. First, given the range of possible values suggested by King, the sign of b should properly be positive, not negative. Recall that, in a conventional specification of the capitalization equation (i.e., with btV on the right-hand side of the equation), the expected sign of b is negative, since an increase in tV should lead to a decrease in V, ceteris paribus. Because $(-btV)$ has essentially been subtracted from both sides of the equation, the sign of b should therefore be positive. Thus, the dependent variable, as redefined, represents the value of the property which would have obtained in the absence of any property tax, but receiving the same level of public services as before. Obviously, this value should be greater than V.

Second, King's equation suggests the capitalization of 1 year's tax bill, whereas the hypothesis would suggest the capitalization of the present discounted value of the stream of future tax bills. Thus, the term btV must be adjusted to reflect the appropriate discount rate, r. These two corrections are incorporated into equation (2), below, which assumes an infinite time horizon:[5]

$$V + \frac{btV}{r} = \beta_0 + \Sigma\beta_i X_i + \Sigma\beta_j P_j. \tag{2}$$

As noted, King suggests using an iterative technique based on observing changes in the level of R^2. However, this technique cannot be legitimately applied to either his equation (1) or the alternative equation (2) without one further adjustment. The difficulty arises because the variance in the dependent variable (and, consequently, the amount of variation to be explained by the independent variables) is increasing as b increases,[6] thus imparting a bias toward findings of low levels of capitalization when the level of R^2 is used as the maximum likelihood criterion.

[5] In private correspondence with the author, King agreed with the validity of these two mathematical corrections. However, he described their omission in the text of his article as a flaw of exposition, noting that his empirical analysis took these two factors into account.

[6] In an analogous case of transformation of the dependent variable (involving a correction for heteroscedasticity), Pindyck and Rubinfeld (1976) observe: "Note that the R^2 measure associated with the weighted least-squares procedure is lower than the R^2 associated with the unweighted procedure *The decline in* R^2 *should not be taken as an indication that the heteroscedasticity correction was incorrect, since the weighted least-squares procedure involves the use of a transformed dependent variable*" (p 102, emphasis added).

To facilitate a valid comparison of the level of R^2 (or, equivalently, of the F-statistic) across equations, the estimating equation may be reexpressed as follows:

$$V = \frac{\beta_0 + \Sigma\beta_i X_i + \Sigma\beta_j P_j}{[1 + (bt/r)]}. \tag{3}$$

With this simple correction, a maximum likelihood estimate of the extent of capitalization may legitimately be obtained by varying (b/r) and successively reestimating, observing changes in the F-statistic.[7]

Estimating Property Tax Capitalization

Because Wallace Oates also made his data available to me, it is possible to compare the results of several alternative specifications of the property tax capitalization equation, including the corrected King specification just proposed. To review, Oates writes his model of property tax capitalization as follows:

$$V = f(T, E, Z, M, R, N, Y, P), \tag{4}$$

where V = median home value by town, T = the effective percentage tax rate, E = annual educational expenditures per pupil, Z = per capita municipal spending on all functions other than local public schools and debt service, M = the linear distance in miles of the community from midtown Manhattan, R = median number of rooms per owner-occupied house, N = percent of houses built since 1950, Y = median family income, and P = percent of families in the community with an annual income of less than $3,000. To these nine variables, King adds a tenth: Tax, the median property tax bill by town, defined as the product of T and V. As Oates and King note, in the single-equation framework just described, the variables E, Z, and T or Tax will be correlated, thereby necessitating the use of two-stage least squares (TSLS) to obtain unbiased estimates of their respective coefficients. Accordingly, all of the equations were estimated using TSLS; the results are presented in table 1.

Table 1 follows the format of King (1977) in presenting the results of the following specifications of the capitalization equation: (Ia) and (IIa) correspond to Oates's (1969) and (1973) estimates, respectively, using the property tax *rate* as the independent variable measuring the extent of capitalization; (Ib) and (IIb) correspond to King's first alternative specification, using the property tax *bill* as an independent

[7] In practice, most standard statistical packages do not print the level of R^2 precisely enough to enable the researcher to identify the maximum likelihood estimate. For this reason, the use of the F-statistic, which conveys the same information as R^2 but in a more precise form, is suggested.

TABLE 1

ALTERNATIVE SPECIFICATIONS OF PROPERTY VALUE EQUATION

Equations	Constant	Log T	Tax	Log E	Log M	R	N	Y	P	Log Z	R^2
Ia	−29	−3 6	.	4 9	−1 3	1 6	06	1 5	3	. . .	.93
	(2 3)	(3.1)		(2 1)	(4 0)	(3 6)	(3.9)	(7.7)	(3 1)		[67.04]
Ib	−38	.	−6 9	5 7	−1.5	1 9	07	1.7	.3	. .	91
	(2.5)		(2 1)	(2 1)†	(3 7)	(3 2)	(4 3)	(7.0)	(2 6)		[49.01]
Ic	−39		−10 8*	5.7	−1 7	2 2	07	1.8	4		‡
	(2.5)			(2 0)	(4.6)	(4 3)	(4 2)	(7 7)	(3 8)		
Id	−44		−17 2*	6.1	−2 0	2 9	07	1.2	.5	. .	95
	(3 0)			(2 3)	(5 5)	(5 7)	(4 7)	(8.8)	(5.1)		[117.95]
IIa	−35	−5 0	. .	4.8	−1 1	1.1	06	1.4	.2	2 7	.93
	(2 7)	(3 8)		(2 1)	(3.4)	(2 2)	(4 2)	(7.1)	(2 4)	(2.2)	[56.07]
IIb	−47	.	−12.0	5 7	−1.5	1.6	08	1 8	3	3 2	.90
	(2 8)		(2.8)	(2.0)	(3 5)	(2 4)	(4.5)	(6.8)	(2 3)	(1 9)	[34.88]
IIc	−46		−11 5*	5 7	−1 5	1.6	08	1.7	3	3 1	‡
	(2 9)			(2.0)	(3 8)	(2.5)	(4.7)	(7.4)	(2 5)	(2 4)	
IId	−60	.	−24 8*	6 1	−1 9	2 2	.09	2.0	4	5.8	97
	(4 2)			(2 4)	(5 3)	(3 8)	(6 1)	(9.3)	(4 1)	(4.6)	[139.49]

NOTE — The variables are defined as in eq (4) Eqq (Ia) and (IIa) are from Oates (1969) and (1973), respectively Eqq (Ib), (Ic) (IIb), and (IIc) are from King (1977) All eqq estimated using 2SLS. T, Tax, E, and Z are endogenous, t-statistics in parentheses, F-statistics in brackets

*Constrained value (see text)

†Incorrectly reported as (3 7) in King (1977)

‡Not reported

variable; while (Ic) and (IIc) are the estimates from King's second, erroneous specification. Finally, (Id) and (IId) are the results using my corrected version of King's specification.

Again following King, computed values for the coefficients of Tax are presented for my equations (Id) and (IId), to facilitate comparisons across equations. These values correspond to capitalization of 100 percent and 145 percent of the property tax bill, respectively.[8] Oates's estimates put capitalization at 66 percent and 92 percent, respectively, while King's estimates from (Ib) and (IIb) indicate 40 percent and 70 percent capitalization and his (erroneous) estimates from (Ic) and (IIc) indicate 63 percent and 67 percent capitalization, respectively. Thus, while King finds that Oates's misspecification of the property tax term results in an overestimation of the extent of property tax capitalization, my results indicate that this misspecification results in *under*estimating the extent of capitalization.

One might reasonably object to the finding of capitalization in excess of 100 percent as inconsistent with capitalization theory. However, the extent of capitalization depends directly on both the discount rate (r) and the time horizon assumed. As figure 1 illustrates, the corrected King technique is only able to identify the *ratio* of b to r which maximizes the value of the F-statistic. Thus, the finding of 145 percent capitalization at a 5 percent discount rate and 40-year time horizon is also consistent with, for example, 100 percent capitalization at a 2.6 percent discount rate and the same time horizon. And, viewed as a homeowner's opportunity cost, r would be defined as the interest rate on mortgages *minus* the rate of appreciation of the house. Given mortgage interest rates prevailing in 1960 (when Oates's data were gathered), therefore, a finding of 100 percent capitalization at a 2.6 percent discount rate may be quite reasonable and not at all inconsistent with the predictions of theory. As further evidence on this point, it is interesting to note that, for the equations in both groups (I) and (II), the highest values of R^2 and of the t-statistics consistently obtain under the corrected King specifications, (Id) and (IId).

Finally, it is quite correct to observe, as did King, that "[i]t can . . . be objected that this equation is inadequate to represent the very complex bundle known as 'housing'" (p. 426). Perhaps, as King speculates, "[t]he instability in the estimated coefficients under different

[8] The estimation technique provides a direct estimate of b, the extent of capitalization; this estimate is then used to compute a coefficient of Tax comparable in interpretation to the coefficients of Oates's original specifications. Eqq. (Id) and (IId) indicate capitalization of 86 percent and 124 percent, respectively, at 5 percent with an infinite time horizon; this is equivalent to capitalization of 100.2 percent and 144.5 percent with a 40-year horizon. The present value of a unit stream for 40 years, discounted at 5 percent, is 17.16. Therefore, $17.16 \times 1.002 = 17.19$ and $17.16 \times 1.445 = 24.80$.

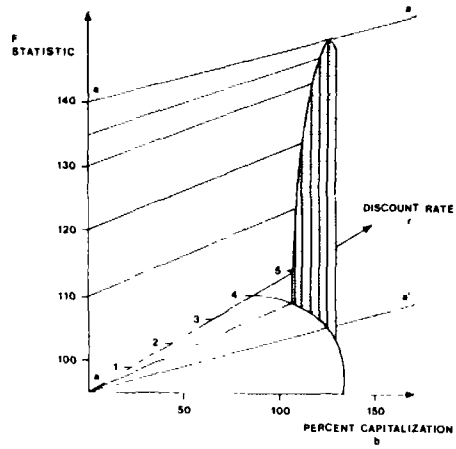


FIG. 1.—Relationship between the F-statistic, discount rate, and the extent of capitalization in the corrected King technique (ray $a'a'$ is a projection of ray aa onto the b-r plane).

specifications noted in the exchange between Pollakowski (1973) and Oates (1973) is very likely attributable to the fact that each included variable acts as a proxy for a great many omitted but correlated variables" (ibid.). To those who would be tempted to cast these results in a similar light, however, I should note that the alternative specifications described herein were tested on a much larger and more detailed data set, with much the same results.

Specifically, the various forms of the capitalization equation were applied to a sample of 1,453 single-family houses sold in San Mateo County, California between June 1, 1969 and June 1, 1970;[9] because observations on suitable instrumental variables were unavailable, all equations were estimated using ordinary least squares. The sample, however, is noteworthy for at least two reasons: First, it consists of observations of actual sales of individual houses, in contrast to Oates's relatively aggregated data. Second, each house is described by a detailed set of physical characteristics, public services received, and tax bills paid (see the Appendix for a full description of the variables included).

[9] The San Mateo County data were originally gathered for a study of capitalization and the Tiebout hypothesis conducted by Jon C. Sonstelie of the University of California, Santa Barbara, and Paul R. Portney of Resources for the Future (Sonstelie and Portney 1980). The author wishes to thank Sonstelie and Portney for kindly making their data available. The results reported here are drawn from a larger study conducted by the author for the Education Commission of the States (Reinhard 1978).

An initial attempt to estimate the extent of capitalization using King's suggested form of the dependent variable (eq. [2], with the appropriate corrections for b and r) found that the value of the F-statistic fell as b was increased; that is, the results implied *no* capitalization of the property tax. Inspection of the difference in the residual sum of squared errors between the equations in which b took on values of 0.2 and 1.0 showed a large difference—an indication of changing variance in the dependent variable.[10]

Estimating the extent of capitalization with equation (3) yielded a maximum value of the F-statistic associated with 100 percent capitalization at a 3.6 percent discount rate and an infinite time horizon.[11] Detailed results of the San Mateo County regression are presented in table A1. Finally, the extent of capitalization was also estimated using a conventional Oates-type equation, with the property tax *rate* included as an independent variable. The results (not shown here) indicated 31 percent capitalization of property taxes, using the same discount rate and time horizon—significantly lower than the estimate obtained with the corrected King specification.

Conclusion

The results reported here do nothing to detract from King's observation that "[m]any studies of property values have used a clearly incorrect specification of the tax effect, one likely to introduce a bias of uncertain direction and magnitude" (p. 430). However, while King's results indicate that the misspecification leads to an upward bias in the tax coefficient for his sample, I find that the use of an Oates-type equation results in a *downward* bias in this coefficient, both when applied to King's data and to data from San Mateo County, California. Further, on the basis of the usual statistical criteria relating to goodness of fit (R^2) and the robustness of individual parameter estimates (t-statistics), my corrected version of King's specification appears to offer a significant improvement over the competing alternative specifications compared herein. Whether this is a general result is, of course, a question for future research. This comment, however, clearly underscores the importance of the choice of estimating equations in influencing the particular results obtained.

[10] This impression was substantiated by the results of the Goldfeld-Quandt test for heteroscedasticity. Test statistic = 1.84, critical value (F with 1,431, 1,431 df $\cong \infty$, ∞ df) = 1.0.

[11] During the 1-year period in which the houses in the sample were sold, the rate on conventional home mortgages was 9 percent. A value of r equal to 3.6 percent implies that the houses were appreciating at a rate of 5.4 percent per year, not an unreasonable assumption for the houses in the sample.

Appendix

Definitions of Variables,* San Mateo County Data

ROOMS	No. of rooms in the dwelling, excluding bathrooms
BATH	No. of bathrooms
QUAL1	Dummy variable indicating average quality of construction
QUAL2	Dummy variable indicating good quality of construction
POOL	Dummy variable indicating the presence of a swimming pool
RMSIZE	Average room size in square feet
AGE	Age of house
LOT	Size of lot in square feet
CULSAC	Dummy variable indicating location on a cul-de-sac or dead-end street
ALLEY	Dummy variable indicating alley behind house
NUIS	Dummy variable indicating presence of a nuisance near property (e.g., bus stop, parking lot, church overflow parking, etc.)
PLAN	Dummy variable indicating well-planned neighborhood
FIRE	Insurance Services Office of California rating of the quality of local fire protection, based on local department, adequacy of water supply, etc. (low rating indicates better protection)
CRIME	FBI Uniform Crime total per 100 residents

* Variables adapted from Sonstelie and Portney (1980)

TABLE A1

REGRESSION RESULTS FOR SAN MATEO COUNTY

Variable	Coefficient	*t*-statistic
CONSTANT	−49,608.65	9.0
ROOMS	4,698.11	13.0
BATH	3,088.41	4.0
QUAL1	977.16	1.1
QUAL2	34,598.91	12.3
POOL	7,978.64	5.2
RMSIZE	137.41	14.6
AGE	−213.36	5.7
LOT	1.142	14.6
CULSAC	3,010.12	2.6
ALLEY	2,099.20	1.1
NUIS	−2,625.10	.5
PLAN	1,141.64	1.1
FIRE	−923.08	1.5
CRIME	−1,945.80	2.7
REC	−430.68	3.1
ROAD	1,580.84	4.8
DIST	−2,794.35	4.1
DIST2	108.59	4.5
AIR	−774.01	3.8
READ	1,468.13	6.2
EXWADA	29.66	6.6

NOTE.—The variables are defined as in the Appendix. Equation estimated using ordinary least squares. All variables except REC are of the predicted sign. Mean of dependent variable = 82,085. $F = 1.997.58$ (22, 1,431 df), $R^2 = .97$.

REC	Recreational expenditures per capita
ROAD	Expenditures per capita on streets and highways
DIST	Linear distance to San Francisco
DIST2	Linear distance to San Francisco, squared
AIR	Number of days per year with at least 0.10 ppm high hour oxidant concentration
READ	Improvement in reading level between first and third grade (measured in months)
EXWADA	Expenditures on elementary and high school education per pupil in average daily attendance (weighted sum)

References

Edel, Matthew, and Sclar, Elliott. "Taxes, Spending, and Property Values: Supply Adjustment in a Tiebout-Oates Model." *J.P.E.* 82, no. 5 (September/October 1974): 941–54.

Epple, Dennis; Zelenitz, Allan; and Visscher, Michael. "A Search for Testable Implications of the Tiebout Hypothesis." *J.P.E.* 86, no. 3 (June 1978): 405–25.

Hamilton, Bruce W. "Capitalization of Intrajurisdictional Differences in Local Tax Prices." *A.E.R.* 66 (December 1976): 743–53. (*a*)

———. "The Effects of Property Taxes and Local Public Spending on Property Values: A Theoretical Comment." *J.P.E.* 84, no. 3 (June 1976): 647–50. (*b*)

King, A. Thomas. "Estimating Property Tax Capitalization: A Critical Comment." *J.P.E.* 85, no. 2 (April 1977): 425–31.

Linneman, Peter. "The Capitalization of Local Taxes: A Note on Specification." *J.P.E.* 86, no. 3 (June 1978): 535–38.

Meadows, George Richard. "Taxes, Spending, and Property Values: A Comment and Further Results." *J.P.E.* 84, no. 4, pt. 1 (August 1976): 869–80.

Oates, Wallace E. "The Effects of Property Taxes and Local Public Spending on Property Values: An Empirical Study of Tax Capitalization and the Tiebout Hypothesis." *J.P.E.* 77, no. 6 (November/December 1969): 957–71.

———. "The Effects of Property Taxes and Local Public Spending on Property Values: A Reply and Yet Further Results." *J.P.E.* 81, no. 4 (July/August 1973): 1004–8.

Pindyck, Robert S., and Rubinfeld, Daniel L. *Econometric Models and Economic Forecasts.* New York: McGraw-Hill, 1976.

Pollakowski, Henry O. "The Effects of Property Taxes and Local Public Spending on Property Values: A Comment and Further Results." *J.P.E.* 81, no. 4 (July/August 1973): 994–1003.

Reinhard, Raymond M. "Capitalization Effects of School Finance Reform: A Guide for California Policymakers." Berkeley: Graduate School of Public Policy, 1978.

Rosen, Harvey S., and Fullerton, David J. "A Note on Local Tax Rates, Public Benefit Levels, and Property Values." *J.P.E.* 85, no. 2 (April 1977): 433–40.

Sonstelie, Jon C., and Portney, Paul R. "Gross Rents and Market Values: Testing the Implications of Tiebout's Hypothesis." *J. Urban Econ.* 7 (January 1980): 102–18.

Tiebout, Charles M. "A Pure Theory of Local Expenditures." *J.P.E.* 64, no. 5 (October 1956): 416–24.

Tax Changes and Cigarette Prices

Michael T. Sumner and Robert Ward

University of Salford

The recurring question of the difference between specific and ad valorem taxes has recently been revived by Barzel (1976), who argued (p. 1185) that ". . . when quality is not fully controlled, the effects of the two types of taxes differ radically. The excise tax induces a substitution of quality for quantity, resulting in a greater increase in the price per unit than predicted by the constant-quality model. The ad valorem tax induces a reduction in quality and lower after-tax price than predicted." Evidence consistent with these hypotheses is found in the response of cigarette prices to tax changes. In the further tests reported by Johnson (1978), prices rise significantly less than the tax in New Hampshire, where the state tax was ad valorem, and significantly more in the remaining states, which impose specific taxes.

The purpose of this note is to question not Barzel's analysis but rather the interpretation of the evidence, on the pedestrian grounds of implausibility. As Barzel himself observes (p. 1194, n. 32), "the size of New Hampshire makes it unlikely that cigarettes would be tailor-made for it. The scope for removing characteristics is quite limited, then." The scope for adding characteristics in other states seems equally limited. To avoid misunderstanding it should be emphasized that the hypothesized variations in quality are a supply-side phenomenon, not a demand-side response to the relative cheapening of superior brands when a specific tax is increased. Johnson's estimates imply that the 9¢ increase in the New Hampshire tax over the period 1954–76 should have reduced the net price by 3.6¢, ceteris paribus; the slightly larger mean tax increase in the remaining 48 continental states should have raised net price by 0.9¢. What changes in the quality of the services sold with cigarettes could have justified such a differential? What are the observable counterparts of the tax changes? In the absence of a convincing answer to those questions, the estimated response of cigarette prices to taxes represents a fact in search of a theory.

[*Journal of Political Economy*, 1981, vol. 89, no. 6]
© 1981 by The University of Chicago 0022-3808/81/8906-0016$01.50

TABLE 1

JOHNSON'S CIGARETTE REGRESSION RESULTS

Variable	Coefficient	Standard Error	T-Value
Constant	13.651	.439	31.078
T_{ij}/Π_i	1.101	.013	87.028

NOTE — R^2 = .965, number of observations = 1,162, 22 year dummies, 50 state dummies, and a New Hampshire tax dummy were also included in the regression equation

The explanation suggested here is that price changes are discontinuous because of the cost of change itself, the uncertainty of competitors' reactions, and the indivisibility of the unit of account. When there is a major change in costs, the retail margins eroded by continuous changes in minor cost components will be restored, so that retail price rises by more than the initiating factor; the latter therefore appears to exert a disproportionate influence on prices, unless appropriate allowance is made for the backlog of small increases which were not worth implementing in isolation. If inflation is not anticipated, the actual price will lag behind the equilibrium price except at the moment of adjustment; if it is anticipated, the actual and equilibrium price levels will coincide on average in the period between adjustments.[1] The tests reported below cannot discriminate between these alternatives but presume that one or other mode of behavior prevailed throughout the period.

To illustrate the relevance of this prosaic suggestion to cigarette prices, consider Johnson's equation:

$$\frac{P_{ij}}{\Pi_i} = \alpha_0 + \alpha_1 \frac{T_{ij}}{\Pi_i} + \text{dummy variables}, \tag{1}$$

where P is cigarette price in cents per pack, T is tax in cents per pack, and Π is the consumer price index on a unit base; subscripts i and j, respectively, denote time and state effects for which the dummy variables are entered as controls. His estimates of the critical parameters are reproduced in table 1. The intercept, which registers the effect on nominal cigarette prices of general inflation, implies that during the first half of the sample period, when the consumer price index was rising very gently, a 1¢ price increase would have occurred at average intervals of more than 5 years in the absence of other disturbances. Tax changes took place more frequently, so it would not be surprising if price changes were concentrated in years when this major component of retail price increased. In the second half of the

[1] We are indebted to Barzel for drawing our attention to this distinction

TABLE 2

CIGARETTE REGRESSION RESULTS: 45 STATES

Variable	Coefficient	Standard Error	T-Value
Constant	−.034	050	670
$\Delta\Pi_t$	8 228	1 270	6.480
ΔT_{ij}	928	036	25.792
$t_{ij}\Sigma_{k=m+1}^{t}\Delta\Pi_k$	2 469	669	3.691
ΔW_t	1.020	111	9.151

NOTE.—R^2 = 744, number of observations = 784

period general inflation accelerated and tax changes became less frequent, so that the modeling of discontinuous price changes becomes less critical.

One simple method of distinguishing between the direct effect of tax changes and their indirect effect of facilitating the restoration of margins is to add an additional variable to a first-differenced nominal version of Johnson's basic formulation. The estimating equation proposed is

$$\Delta P_{ij} = \alpha\Delta\Pi_t + \beta\Delta T_{ij} + \gamma t_{ij} \sum_{k=m+1}^{t} \Delta\Pi_k, \tag{2}$$

where $t_{ij} = 1$ if $\Delta T_{ij} \neq 0$, $= 0$ otherwise. Subscript m indicates the date of the previous tax change in state j. This formulation differs from a transformed version of Johnson's equation (1) in not permitting the effect of general inflation to differ across states and time through the dummy variables; instead, effects of time specific to the cigarette industry were captured by adding the change in wholesale cigarette prices (ΔW) to equation (2), and the residuals were scrutinized for state effects.

Thus augmented, equation (2) was estimated for 45 states which impose a specific tax[2] with the results shown in table 2. The intercept is appropriately insignificant and, as anticipated, wholesale price changes have a coefficient which does not differ significantly from unity. The inclusion of the latter variable reduces the estimated effect of general inflation but does not affect the main conclusions. These are that the diffused sources of price changes represented by general inflation enter both directly and in conjunction with tax changes; and that once allowance is made for the indirect influence of tax changes in effecting backlogged price increases, the coefficient on tax change

[2] We are grateful to Johnson for providing his data. First differencing and, more importantly, the construction of the backlog variable reduced the sample from 1,056 to 784 observations and eliminated three states in which only a single tax change occurred

TABLE 3

CIGARETTE REGRESSION RESULTS: 11 STATES

Variable	Coefficient	Standard Error	T-Value
Constant	.046	.081	.569
$\Delta\Pi_t$	5.158	2.252	2.291
ΔT_0	.698	.061	11.442
${}_{t0}\Sigma_{k=m+1}\Delta\Pi_k$	5.910	1.230	4.805
ΔW_t	1.236	.196	6.300

NOTE.—R^2 = .727, number of observations = 231

itself becomes significantly less than unity. Finally, inspection of the residuals does not indicate any need to incorporate dummy variables for state effects: The largest mean state error is less than half the standard deviation of the residuals.

It was suggested above that the association between tax changes and the restoration of margins eroded by mild inflation would be stronger the lower the general inflation rate. That suggestion was verified by fitting the estimating equation to the sample of 11 states for which data begin in 1956, the earliest possible date. The major differences in results are immediately apparent in table 3. The most obvious change is the lower point estimate of the tax-change coefficient. Second, the sum of the coefficients on the catch-up variable and contemporaneous price change remains about the same, but their relative size changes dramatically: Tax changes are much more important as an occasion for effecting backlogged price increases in this sample, in which the earlier, more quiescent inflation experience is more heavily weighted.

The major question remaining is how well the model fits New Hampshire. For the period 1955–76[3] the mean forecast change exceeds the actual increase by 13 percent; Theil's U-statistic is 0.362. These figures are difficult to appraise in isolation but do not seem out of line with normal forecasting standards. That impression is strengthened by the observation that in 10 states the mean residual exceeds the average forecast error in New Hampshire. Finally, the largest individual error occurs in 1975, when the predicted price increase exceeded the actual rise by more than 1.5¢ to yield a proportionate error of 50 percent. This is particularly inconvenient for

[3] To avoid losing early observations additional data were collected from the Tobacco Tax Council (1978). The tax series for New Hampshire differs from that provided by Johnson in two respects: There is an automatic change in 1957 which does not appear in Johnson's data until 1960; and a single legislated change in 1971 is split between 1971 and 1972 in Johnson's data. The revised data are used here. The analysis of the forecasts reveals no difference between the impact of legislated and automatic tax changes.

Barzel and Johnson because the form, as well as the rate, of tax was changed in that year from ad valorem to specific.

For present purposes, the most important result is the negative one that the tax-change coefficient is not significantly greater than unity; but, from a broader perspective, the conclusion that it is significantly less than unity is at least equally interesting. A possible explanation for the difference between this coefficient and that on wholesale price changes is interstate competition. In this context it is particularly noteworthy that the state which lies furthest below the regression plane is geographically isolated Maine, where the levels of taxes and prices have consistently exceeded those in neighboring New Hampshire. Of the two price increases not associated with intrastate tax changes and exceeding 1¢ in magnitude, one coincided with and the other followed a year after tax changes in New Hampshire. Massachusetts and Vermont also lie below the regression plane and have consistently experienced higher tax and price levels than New Hampshire; in both cases, three of four large price changes unconnected with domestic tax changes coincided with tax changes in New Hampshire. These selective observations are suggestive but hardly constitute evidence. It is, however, beyond the scope of this note to construct a scalar measure of state openness and to test its interaction with tax changes.

There remain unexplored possibilities of modeling discontinuous price changes more faithfully, but for present purposes it is not necessary to pursue them. It has been shown that there is no difference between the effects of specific and ad valorem taxes on cigarette prices and that the impact of taxes per se on prices is significantly less than one for one.

References

Barzel, Yoram. "An Alternative Approach to the Analysis of Taxation." *J P E* 84, no. 6 (December 1976): 1177–97.

Johnson, Terry R. "Additional Evidence on the Effect of Alternative Taxes on Cigarette Prices." *J.P E.* 86, no. 2, pt. 1 (April 1978): 325–28.

Tobacco Tax Council, Inc. *The Tax Burden on Tobacco: Historical Compilation.* Vol. 13. Richmond, Va.: Tobacco Tax Council, 1978

Miscellany

'Twas a Night in the Sixties

'Twas a night in the sixties
And all through the land
Unemployment was falling
Inflation in hand.

The stock market was rising,
Without any care,
In hopes a Dow thousand
Soon would be there.

The Keynesians were snuggled
Secure in their Chairs,
While visions of multipliers
Allayed all their cares.

Paul with his textbook
And Art with his gap
Had settled their brains
For a long postwar nap.

When out in the land
There arose such a clatter,
A voice that was crying
That money could matter.

This study extends earlier work by C. C. Moore, to whom my debt is obvious. It was presented at the 1980 Harvard Economics Department Student-Faculty Christmas party. Although I am always grateful to the National Science Foundation and National Bureau of Economic Research, neither organization provided any financial support for this project.

[*Journal of Political Economy*, 1981, vol. 89, no. 6]
© 1981 by The University of Chicago. 0022-3808/81/8906-0017$01.50

Away from their desks
They flew in a flash
To see who was claiming
Such power for cash.

They looked at their models
With equations precise,
That gave semblance of proof
To conclusions so nice.

When what to their wondering
Eyes should appear
But a miniature sleigh
With eight tiny reindeer

With a little old driver
Who was having such fun
They knew in a moment
It must be Milton.

More numerous than eagles
His supporters they came
And he whistled and shouted
And called them by name.

First John Say and then Hume,
Then Marshall and Mill,
Now Brunner and Meltzer
And Anna and Phil.

From the U. of Chicago
To Minneapolis–St. Paul
Then dash away! Dash away!
Dash away all!

As economic theories
With which economists play
When they meet with an obstacle
Assume it away,

So off to the journals,
Their papers they flew,
With monetarist theorems,
Rational expectations too.

And even in Cambridge
Was heard the new truth,
The theorems and lemmas
Of each little proof.

The Keynesian thinkers
Were spinning around
When onto the scene,
Milton came with a bound.

He was dressed all in gold
From his head to his foot
And his ideas were polished
And ready to put.

"Velocity's stable,
*M*1 and *M*2,
Which shows what the Fed
Shouldn't be trying to do."

"That curve by Phillips
It really is straight
And the cost of funds
Is the *real* interest rate."

He wrote many a word,
And with evidence too.
At the NBER
His volumes they grew.

His ideas how simple.
He puts them so well.
It would be no wonder
When he got his Nobel.

A wink of his eye
And a nod of his head
Soon gave Keynesians to know
They had something to dread.

Then he sprang to his sleigh
To his team gave a whistle
And away they all flew
Like the down of a thistle.

But I heard him exclaim
As he drove out of sight,
"Keep freedom for all,
And keep money tight."

MARTIN S. FELDSTEIN

National Bureau of Economic Research
Harvard University

Policy Research

The title of this discourse is "Policy Research," a contradiction in terms. Its purpose is to explain that the demise of this once proud office was not an accident of circumstance but an inexorable consequence of a tendency which has been at the heart of political systems since the Middle Ages.

The elements which come together to explain this phenomenon were collected in various places.

The first element appeared at the International Trade Commission, where for a year I was privileged to observe both the deliberations of the commission to reach decisions and the diligent efforts of their staff to produce a staff report. But there was no relationship of one to the other. Indeed, the place where the commission's decision and the staff report came together was always the printshop. I notice that though its purpose was obscure, the staff report did serve a function—to misdirect and to obfuscate. Having been exposed to the institutionalist tradition, I was confident that this function served some purpose, but I was unable to determine what that purpose was.

Years passed, and I observed that in nearly every town and city of Europe, there was a "Jews quarter" or *Ruelle des Juifs* in the old commercial part of town. This treasure, too, I stored in the backyard of my mind.

Finally, a book by Lester Thurow reminded us that zero-sum decisions are *very* difficult for a democratic political process to make. Then things began to fall into place. The purpose of misdirection and obfuscation became obvious. By making the decision process difficult to understand, they discourage those with a diffuse or marginal interest from participating in the decision; that is, they tend to disenfranchise the opponents to a petition for protection and hence transform a zero-sum *situation* into a positive-sum *decision*—a decision from which all the *effective* participants will gain.

This, then, is the essence of the decision process. Unless the great clamor of procedure confuses one side about what is going on, there can be no decision, that is, no policy.

Presented September 4, 1980, at the U.S. Treasury Department, on the occasion of the departure of Finger and the dissolution of the Office of Trade Research, of which Finger was the director.

[*Journal of Political Economy*, 1981, vol. 89, no. 6]
© 1981 by The University of Chicago 0022-3808/81/8906-0018$01.50

Of course, if procedure confuses both sides, the government is in the ideal situation of not having to make a decision at all. This occurs with some frequency.

But if one side sees through the obfuscation and the other does not (the result which the system seems designed to produce), then the aware side's interests will be disproportionately weighted in the decision process, and equilibrium will be "too far" in their direction. In summary, the policy process in an open political system

(1) does not do a lot of things it should do,
(2) appears to do a lot of things it does not do, and
(3) overdoes everything it does do.

Good research which accurately identifies the costs and benefits of a course of action is, in any of these avenues, more likely to be embarrassing than welcome. It returns attention to the direct conflict which had to be *avoided* before a course of action—a policy—could be set. This means that a person who delivers a novel piece of analytical work to his Political Master will be made to feel like a woman who brings a new child to her husband—who had a vasectomy 5 years before. The Master is less likely to share her joy in the birth of a new thing than to decry this evidence of her having conspired with mean spirits. Political responsibility is, you see, the ultimate intellectual vasectomy.

The history of research in government agencies is like the history of Jews in pre–Industrial Revolution Europe. The Gentiles of a town would decide that they wanted to establish their town as a center of commerce. They would then invite the practitioners of this science into their town and set aside a place where they might live and practice. But, the Jews would practice their trade by rules so basic that they would not allow their gentlemen competitors, to whom commerce was a diversion and a livelihood but not a life-style, to turn a profit. So, slowly the number of streets open to the Jews would be restricted and the number of products they were allowed to trade reduced. Or, they would be limited to the wholesale trade, in the hope that gentlemen could turn a profit retailing the merchandise the Jews gathered.

But their numbers would dwindle, not so much because they had been sent away but because other Gentiles in other towns had decided that they should establish centers of commerce and had lured them away, to begin the process again.

As we see across Europe the signs announcing the former presence of Jews, we see across organization charts of government agencies the street signs of research. But the original citizens of these quarters have usually gone, and their places have been taken by reform congregations, or even Gentiles. Thus, we, the researchers, like the Jews of the Middle Ages, are cursed to wander from agency to agency,

deprived of home and property by our inability to lose track of what we do.[1]

J. MICHAEL FINGER

World Bank

[1] The mortal remains (former employees) of the dissolved Office of Trade Research were sent away with a memento inscribed "There are no Statesmen in this business. Trade theory is about whose hand is in whose pocket and trade policy is about who should take it out." It was noted gratefully that as they departed, their temple would be dismantled and the signs in their quarter taken down. Thus they would be spared the indignity of seeing a reform congregation move in and eat pork and write briefing papers where once they followed the old law. The spirit of the office was also dispatched. It was represented by a full frontal picture of a charging infantry brigade, inscribed, "We owe our success to a simple underlying philosophy."

Journal of Political Economy

EDITED BY JACOB A. FRENKEL, ROBERT E. LUCAS, JR.,
SAM PELTZMAN, AND GEORGE J. STIGLER
IN CO-OPERATION WITH OTHER MEMBERS OF THE
DEPARTMENT OF ECONOMICS AND THE
GRADUATE SCHOOL OF BUSINESS OF
THE UNIVERSITY OF CHICAGO
AND OUTSIDE REFEREES

Volume 89

February–December 1981

THE UNIVERSITY OF CHICAGO PRESS
CHICAGO, ILLINOIS

© 1981 BY THE UNIVERSITY OF CHICAGO ALL RIGHTS RESERVED
PUBLISHED FEBRUARY,
APRIL, JUNE, AUGUST, OCTOBER,
DECEMBER 1981
BY THE UNIVERSITY OF CHICAGO PRESS

Index to Volume 89

Articles and Notes

Abel, Andrew B. Taxes, Inflation, and the Durability of Capital 548
Allen, Robert C. Accounting for Price Changes: American Steel Rails, 1879–1910 512
Anderson, Ronald W., and Danthine, Jean-Pierre Cross Hedging ... 1182
Appelbaum, Elie, and Katz, Eliakim. Market Constraints as a Rationale for the Friedman-Savage Utility Function 819
Bailey, Martin J. See *Olson, Mancur*
Barro, Robert J. Output Effects of Government Purchases 1086
Bassett, Gilbert W., Jr. Point Spreads versus Odds 752
Battalio, Raymond C.; Kagel, John H.; Rachlin, Howard; and Green, Leonard. Commodity-Choice Behavior with Pigeons as Subjects 67
Baumol, William J., and Wolff, Edward N. Subsidies to New Energy Sources: Do They Add to Energy Stocks? 891
Bhagwati, Jagdish N. See *Brecher, Richard A.*
Blejer, Mario I., and Leiderman, Leonardo. A Monetary Approach to the Crawling-Peg System: Theory and Evidence 132
Blinder, Alan S. Temporary Income Taxes and Consumer Spending 26
Block, Michael Kent; Nold, Frederick Carl, and Sidak, Joseph Gregory. The Deterrent Effect of Antitrust Enforcement 429
Bolnick, Bruce R. Recalculating the Scientific Tariff 192
Boyer, Kenneth D. Equalizing Discrimination and Cartel Pricing in Transport Rate Regulation 270
Brecher, Richard A., and Bhagwati, Jagdish N. Foreign Ownership and the Theory of Trade and Welfare 497
Buiter, Willem H. Time Preference and International Lending and Borrowing in an Overlapping-Generations Model 769
Burns, Michael E., and Walsh, Cliff. Market Provision of Price-excludable Public Goods: A General Analysis 166
Camacho, Antonio, and White, William D. A Note on Loss of Control and the Optimum Size of the Firm 407
Cardoso, Eliana A. The Great Depression and Commodity-exporting LDCs: The Case of Brazil 1239
Cassing, James. On the Relationship between Commodity Price Changes and Factor Owners' Real Positions 593
Chamley, Christophe. The Welfare Cost of Capital Income Taxation in a Growing Economy 468
Cohen, Kalman J., Maier, Steven F., Schwartz, Robert A., and Whitcomb, David K. Transaction Costs, Order Placement Strategy, and Existence of the Bid-Ask Spread 287
Daly, Vince, and Hadjimatheou, George. Stochastic Implications of the Life Cycle–Permanent Income Hypothesis: Evidence for the U.K. Economy 596
Daniel, Betty C. International Transmission of a Real Shock under Flexible Exchange Rates: A Comment 813

Danthine, Jean-Pierre. See *Anderson, Ronald W.*
Davies, James B. Uncertain Lifetime, Consumption, and Dissaving in Retirement 561
Dertouzos, James N., and Pencavel, John H. Wage and Employment Determination under Trade Unionism: The International Typographical Union 1162
Diamond, Peter A. Mobility Costs, Frictional Unemployment, and Efficiency 798
Driskill, Robert A. Exchange-Rate Dynamics: An Empirical Investigation 357
Eckard, E. Woodrow, Jr. Concentration Changes and Inflation: Some Evidence 1044
Epple, Dennis, and Zelenitz, Allan. The Implications of Competition among Jurisdictions: Does Tiebout Need Politics? 1197
Feldstein, Martin S. 'Twas a Night in the Sixties 1266
Finger, J. Michael. Policy Research 1270
Flavin, Marjorie A. The Adjustment of Consumption to Changing Expectations about Future Income 974
Friedman, David. Why There Are No Risk Preferrers 600
Frenkel, Jacob A. Flexible Exchange Rates, Prices, and the Role of "News": Lessons from the 1970s 665
Galenson, David W. The Market Evaluation of Human Capital: The Case of Indentured Servitude 446
Green, Leonard. See *Battalio, Raymond C.*
Hadjimatheou, George. See *Daly, Vince.*
Helpman, Elhanan. An Exploration in the Theory of Exchange-Rate Regimes 865
Hercowitz, Zvi. Money and the Dispersion of Relative Prices 328
Hey, John D., and McKenna, Chris J. Consumer Search with Uncertain Product Quality 54
Howitt, Peter. Activist Monetary Policy under Rational Expectations . 249
Jaeger, Elizabeth. To Save or Savor: The Rate of Return to Storing Wine 584
Johnsen, Christopher. See *Weingast, Barry R.*
Kagel, John H. See *Battalio, Raymond C.*
Katz, Eliakim. See *Appelbaum, Elie.*
Klein, Benjamin, and Leffler, Keith B. The Role of Market Forces in Assuring Contractual Performance 615
Kopcke, Richard W. Inflation, Corporate Income Taxation, and the Demand for Capital Assets 122
Kotlikoff, Laurence J., and Spivak, Avia. The Family as an Incomplete Annuities Market 372
Kotlikoff, Laurence J., and Summers, Laurence H. The Role of Intergenerational Transfers in Aggregate Capital Accumulation 706
Krugman, Paul R. Intraindustry Specialization and the Gains from Trade 959
Landsburg, Steven E. Taste Change in the United Kingdom, 1900–1955 92
Laumas, G. S. Discount Rate and Wealth 196
Lazear, Edward P., and Rosen, Sherwin. Rank-Order Tournaments as Optimum Labor Contracts 841
Leiderman, Leonardo. See *Blejer, Mario I.*
Leffler, Keith B. See *Klein, Benjamin.*
Liviatan, Nissan. Monetary Expansion and Real Exchange Rate Dynamics 1218
McKenna, Chris J. See *Hey, John D.*

MaCurdy, Thomas E. An Empirical Model of Labor Supply in a Life-Cycle Setting 1059
Maier, Steven F. See *Cohen, Kalman J.*
Meltzer, Allan H., and Richard, Scott F. A Rational Theory of the Size of Government 915
Mumy, Gene E. A Superior Solution to Captain MacWhirr's Problem: An Illustration of Information Problems and Entitlement Structures .. 1039
Nold, Frederick Carl. See *Block, Michael Kent.*
Obstfeld, Maurice. Macroeconomic Policy, Exchange-Rate Dynamics, and Optimal Asset Accumulation 1142
Olson, Mancur, and Bailey, Martin J. Positive Time Preference 1
Pencavel, John H. See *Dertouzos, James N.*
Rachlin, Howard. See *Battalio, Raymond C.*
Ray, Edward John. The Determinants of Tariff and Nontariff Trade Restrictions in the United States 105
Reinhard, Raymond M. Estimating Property Tax Capitalization: A Further Comment 1251
Richard, Scott F. See *Meltzer, Allan H.*
Rose-Ackerman, Susan. Does Federalism Matter? Political Choice in a Federal Republic 152
Rosen, Sherwin. See *Lazear, Edward P.*
Sargent, Thomas J. Interpreting Economic Time Series 213
Schmalensee, Richard. Economies of Scale and Barriers to Entry 1228
Schwartz, Robert A. See *Cohen, Kalman J.*
Shefrin, H. M. See *Thaler, Richard H.*
Shepsle, Kenneth A. See *Weingast, Barry R.*
Sidak, Joseph Gregory. See *Block, Michael Kent.*
Sims, Christopher A. What Kind of a Science Is Economics? A Review Article on *Causality in Economics* by John R. Hicks 578
Spivak, Avia. See *Kotlikoff, Laurence J.*
Starrett, David A. Land Value Capitalization in Local Public Finance . 306
Stuart, Charles E. Swedish Tax Rates, Labor Supply, and Tax Revenues 1020
Summers, Lawrence H. See *Kotlikoff, Laurence J.*
Sumner, Daniel A. Measurement of Monopoly Behavior: An Application to the Cigarette Industry 1010
Sumner, Michael T., and Ward, Robert. Tax Changes and Cigarette Prices . 1261
Tauchen, George E. Some Evidence on Cross-Sector Effects of the Minimum Wage 529
Thaler, Richard H., and Shefrin, H. M. An Economic Theory of Self-Control 392
Tomes, Nigel. The Family, Inheritance, and the Intergenerational Transmission of Inequality 928
Walsh, Cliff. See *Burns, Michael E.*
Ward, Robert. See *Sumner, Michael T.*
Weber, Warren E. Output Variability under Monetary Policy and Exchange Rate Rules 733
Weber, Warren E. Prior Information and the Observational Equivalence Problem 411
Weingast, Barry R.; Shepsle, Kenneth A.; and Johnsen, Christopher. The Political Economy of Benefits and Costs: A Neoclassical Approach to Distributive Politics 642
Whitcomb, David K. See *Cohen, Kalman J.*
White, William D. See *Camacho, Antonio.*

Wilde, Louis L. Information Costs, Duration of Search, and Turnover: Theory and Applications 1122
Wolff, Edward N. See *Baumol, William*
Zelenitz, Allan. See *Epple, Dennis.*

Accession Number. 82790
Date 30.1.83

Reviews

Bhagwati, Jagdish N. *Foreign Trade Regimes and Economic Development: Anatomy and Consequences of Exchange Control Regimes.* By Deepak Lal 826
Courchene, Thomas J. *Money, Inflation, and the Bank of Canada: An Analysis of Canadian Monetary Policy from 1970 to Early 1975.* By Malcolm Knight 199
Easterlin, Richard A., ed. *Population and Economic Change in Developing Countries.* By John C. Caldwell 830
Eckert, Ross D. *The Enclosure of Ocean Resources: Economics and the Law of the Sea.* By Anthony Scott 416
Edwards, Franklin R., ed. *Issues in Financial Regulation.* By P. Michael Laub 417
Ehrenberg, Ronald G. *The Regulatory Process and Labor Earnings.* By Orley Ashenfelter 601
Friedman, Benjamin M., ed. *New Challenges to the Role of Profit.* By Gerald P. O'Driscoll, Jr. 421
Hayek, Friedrich A. *Law, Legislation and Liberty: A New Statement of the Liberal Principles of Justice and Political Economy. Vol. 3: The Political Order of a Free People.* By Clas Wihlborg 603
Hicks, Sir John R. *The Crisis in Keynesian Economics.* By Karl Brunner 1052
Jacobsson, Erin E. *A Life for Sound Money: Per Jacobsson. His Biography.* By Alec Cairncross 202
Lippman, Steven A., and McCall, John J., eds. *Studies in the Economics of Search.* By Dale T. Mortensen 609
Lydall, Harold. *A Theory of Income Distribution.* By Martin Bronfenbrenner 611
MacAvoy, Paul W. *The Regulated Industries and the Economy.* By Wesley A. Magat 834
Macfarlane, Alan. *The Origins of English Individualism: The Family, Property and Social Transition.* By Donald N. McCloskey 837
Patinkin, Don. *Keynes' Monetary Thought: A Study of Its Development.* By James Tobin 204
Phelps, Edmund S. *Studies in Macroeconomic Theory: Volume 1: Employment and Inflation.* By Herschel I. Grossman 207
Popkin, Samuel L. *The Rational Peasant: The Political Economy of Rural Society in Vietnam.* By Donald N. McCloskey 837
Repetto, Robert. *Economic Equality and Fertility in Developing Countries.* By Carmel U. Chiswick 423
Throsby, C. D., and Withers, G. A. *The Economics of the Performing Arts.* By William J. Baumol and Hilda Baumol 425
Thurow, Lester C. *The Zero-Sum Society: Distribution and the Possibilities for Economic Change.* By Wilfred Beckerman 1054